Algorithmic fairness auditing: can discrimination by algorithms be prevented?

Final thesis for the master Artificial Intelligence at Utrecht University

Author:	Lukas Jan Graff
Student number:	6557236
Supervisor:	Dr. Mirko Tobias Schäfer
Second supervisor:	Prof. dr. Albert Ali Salah
Final submission:	20-03-2024

Preface

This thesis is about reducing the risk of discrimination by algorithms in practice. It is aimed at any person who is involved in, wants to be involved in or is otherwise interested in using algorithms in a way that is fair or ethical, regardless of the academic or professional background of this person. Hence, this thesis does not assume background knowledge of computer science, non-discrimination law or any other expertise involved in making algorithms fair. Although the research in this thesis is directed at the Netherlands specifically, many of its themes and findings are expected to be of a more universal nature and applicable to other countries as well, especially other EU member states that share an important body of non-discrimination legislation.

For me personally, first and foremost, this thesis is the final step in completing my master's degree in Artificial Intelligence (AI). It shows what I have learnt during more than five years (Bachelor and Master combined) of studying this subject at Utrecht University (UU). To me, the beauty of AI is in the countless disciplines it connects, and this versatility was also very prominent in the way the subject is approached by the UU. During my time there I had courses on programming, machine learning, philosophy, cognitive science, mathematics, logic, linguistics, humanities, applied psychology, science communication and even close reading of prose and poetry (although I must admit that was mostly for fun and out of personal interest). I am very thankful to the UU for allowing me to explore so many sides of AI and of myself. To stay in tune, this thesis will explore many sides of AI as well, by connecting computer science to social science and law.

Not only is this thesis the final product of my master's degree, but it is also the final lesson learnt in it. The interdisciplinarity that is key to the topic of my thesis, required me to familiarise myself with scientific disciplines outside of my academic comfort zone and use scientific methods I had never used before. During my thesis project, I truly learned to see my area of expertise in perspective and learned that it should never be considered in isolation. By doing a research project that was larger than any project I have ever carried out before, I also learned valuable lessons about structure, self-motivation and discipline. And as any good research might befit, now that it is finished, I feel that if I were to start all over again, I would approach everything very differently. This does not detract from my thesis as it is now, because in the end all worked out. It merely serves as a testament to show how much this project made me learn.

I would like to thank Utrecht University's Data School for allowing me to join their sessions in which they used their acclaimed methods for discussing data ethics and assessing the impact of algorithms on human rights. Even though we eventually discovered that joining these sessions would not provide me the data I needed for my research, they helped me a lot in familiarising myself with algorithm ethics in practice. Special thanks go to Jeroen Bakker, who was my daily supervisor during my time at the Data School and has been very involved, and to Mirko Schäfer, the cofounder of the Data School and associate professor at the UU, who did not hesitate to become first supervisor for my thesis project after I had struggled to find one for quite some time.

Content

List of tables and figures

Tables:

Figures:

List of acronyms

ACM FAccT	ACM Conference on Fairness, Accountability, and Transparency
AI	Artificial Intelligence
AWGB	The Dutch Equal Treatment Act (Netherlands) (Dutch: *Algemene wet gelijke behandeling*)
CRM	The Netherlands Institute for Human Rights (Dutch: *College voor de Rechten van de Mens*)
DIR (in formulas)	Disparate Impact Ratio
EC	European Commission
ECJ	European Court of Justice
EU	European Union
FNR	False Negative Rate
FPR	False Positive Rate
FRAIA	Fundamental Rights and Algorithms Impact Assessment (Dutch: *Impact Assessment Mensenrechten en Algoritmes (IAMA)*)
GDPR	General Data Protection Regulation (EU)
GW	Constitution of the Netherlands (*Dutch: Grondwet*)
ISO	International Organization for Standardization
ML	Machine Learning
NPV	Negative Predictive Value
PPV	Positive Predictive Value
QCA	Qualitative Comparative Analysis
STS	Science and Technology Studies
SyRi	System (for) Risk Indication (Dutch: *Systeem Risico Indicatie*)
USA	United States of America
WCRM	Netherlands Institute for Human Rights Act (Netherlands) (Dutch: *Wet College voor de rechten van de mens*)
WGBH/CZ	The Equal Treatment (Disabled and Chronically Ill People) Act (Netherlands) (Dutch: *Wet gelijke behandeling op grond van handicap of chronische ziekte*)
WGBLA	The Equal Treatment in Employment (Age Discrimination) Act (Netherlands) (Dutch: *Wet gelijke behandeling op grond van leeftijd bij de arbeid*)
WGBMV	The Equal Treatment (Men and Women) Act (Netherlands) (Dutch: *Wet gelijke behandeling van mannen en vrouwen*)
WOA	The Working Hours Discrimination Act (Netherlands) (Dutch (informally): *Wet onderscheid arbeidsduur*)
WOBOT	The Definite and Indefinite Duration Discrimination Act (Netherlands) (Dutch (informally): *Wet onderscheid bepaalde en onbepaalde tijd*)

Abstract

Algorithms are increasingly used in decision-making. As numerous scandals show, this introduces new risks of discrimination on large scales. Algorithmic fairness audits have often been proposed as a binding method to reduce this and other risks. Hence, this research investigates what role auditing can play in ensuring algorithmic fairness, in terms of non-discrimination. Strictly defined, auditing leaves no room for subjective interpretation, meaning that all choices faced when assessing algorithmic fairness should be eliminated to create an audit framework. Hence our research focusses on detecting and eliminating these choices. Firstly, we identify the normative choices that are faced when assessing algorithmic fairness from a computer science perspective. Secondly, we investigate to what extent Dutch non-discrimination legislation prescribes how these choices should be made. We discover that some important, normative choices are left open by law. Hence, thirdly, we use informal conversations with algorithmic fairness practitioners to explore best practices in algorithmic fairness assessments to find alternative ways of deciding on these normative choices. Finally, we conclude that algorithmic fairness audits cannot be used directly to ensure non-discrimination. However, both internal and external audits can have a more indirect use in ensuring non-discrimination by ensuring the soundness of either the procedure of internal algorithmic fairness assessments or the documentation thereof.

Introduction

As a student in Artificial Intelligence (AI) with a great interest in social issues and social justice, I have always been mindful of the potential downsides of using AI. Striking examples of an algorithm (the COMPAS algorithm) used by USA judges to inform their decisions, that seemingly to (re)produce racist prejudices (Larson et al., 2016) or a hiring algorithm used by tech giant Amazon that learnt to be sexist (Dastin, 2018), quickly made me realise that discriminatory algorithms are a serious issue. In the following years, a series of scandals involving discriminatory algorithms used in the Netherlands public sector, quickly showed that the phenomenon of algorithmic discrimination was not something that only affected foreign countries.

The *Dutch childcare benefits scandal*[1] is one of the most infamous examples of such a scandal. This scandal was caused by failing Dutch government policy aimed at detecting welfare fraud and reclaiming the money that parents had supposedly illegitimately received. However, this policy led to tens of thousands of wrongful accusations and serious financial distress for many parents (Geiger, 2021; Henley, 2021; NOS Nieuws, 2024). When a report revealing the full extent of the scandal was published, this even led to the formal resignation of the then government (albeit two months before their term would have ended anyway.) (Amaro, 2021; NOS Nieuws, 2021) For a long time, an algorithm (a risk classification model) has been used in selecting parents for suspicion of fraud. Both the Netherlands Institute for Human Rights[2] and the Dutch Data Protection Authority[3] have judged this algorithm to be discriminatory against people of non-Dutch descent (Autoriteit Persoonsgegevens, 2020; College voor de Rechten van de Mens, 2022; *verdict Belastingdienst/Toeslagen*)

The Dutch government also faced serious criticism over the use of an algorithmic model used in detecting welfare fraud, known as *SyRI* (System (for) Risk Indication).[4] Eventually, the use of SyRI was forbidden by Dutch court, because it violated human rights to an extent that was not proportional for the purpose it served. More specifically, the court established an infringement of the right to privacy as defined by the European Convention on Human Rights, but the court also stated that algorithmic risk models like SyRI were at risk of having discriminatory effects on citizens. The lack of

Relevance for AI

As this introduction shows, algorithmic discrimination is a pressing issue. Even simple algorithms, which arguably cannot be called (artificially) intelligent, can and often do cause discrimination (e.g. the risk classification model used in the Dutch childcare benefits scandal). Yet, the use of the common AI technique (deep) machine learning introduces new problems for detecting this discrimination, since it may be hard to understand the logic according to which algorithms using this technique operate and hence it may be hard to understand whether these algorithms discriminate.

This is why most of the literature on algorithmic fairness focusses on machine learning algorithms specifically. The fairness metrics proposed in this literature are suited for machine learning algorithm specifically because they never assume any (possibility) of understanding of the logic followed by the algorithm. Furthermore, these metrics are often based on mathematical/statistical concepts that are important in the field of AI anyway.

Hence, although the problems and solutions discussed in this thesis might not be unique to AI/machine learning algorithms, they are incredibly relevant to the field of AI and will be increasingly relevant as AI techniques will become more mainstream in decision-making algorithms.

[1] Dutch: *toeslagenaffaire* or *toeslagenschandaal*.

[2] Dutch: *College voor de Rechten van de* Mens; More on this institute and this specific verdict in chapter 2.

[3] Dutch: *Autoriteit Persoonsgegevens*

[4] Dutch: *Systeem Risico Indicatie*

transparency provided by the government about the use and functioning of SyRI was considered especially bad, because it diminished the ability to control for the presence of potential discriminatory effects (College voor de Rechten van de Mens, 2020; Staat der Nederlanden; van Bekkum & Borgesius, 2021). At a local level, the Dutch municipality of Rotterdam used a similar risk model for detecting welfare fraud. This algorithm was trained using machine learning. The Rotterdam Court of Audit[5] has pointed out that the municipality of Rotterdam had failed to test for indirect discrimination by the algorithm, even though its inclusion of *fluency in the Dutch language* as input feature led to indirect discrimination based on race or nationality (Rekenkamer Rotterdam, 2021).[6] Furthermore, a thorough journalistic investigation of this algorithm did, in fact, show that this variable did greatly impact the outcome of the algorithm and that the algorithm also showed bias against women (Geiger et al., 2023). Rotterdam stopped using this algorithm in 2021. Until 2023 the Dutch executive government body for education used an algorithm to detect fraud in receiving student grants. A collaboration of Dutch news and journalistic research organisations found out that the indicators for selection by this algorithm were strongly associated with certain ethnic backgrounds and that 97 percent of the formal objections against allegations of student grant fraud were made by people with a migration background (Belleman et al., 2023; Ersoy & van der Gaag, 2023). These findings raised serious suspicion of discrimination leading the responsible minister to terminate the use of the algorithm until its potential discriminatory effects would be investigated (van der Gaag & Ersoy, 2023).

As this list of scandals shows, algorithmic discrimination is a real and pressing issue in the Netherlands. The research field that aims to detect and combat this form of discrimination, is often called algorithmic fairness. This thesis will investigate how algorithmic fairness, in terms of non-discrimination, can be improved. Given my own familiarity with the Netherlands and the history of algorithmic discrimination in that country, this investigation will be situated in context of the Netherlands. In response to the reported algorithmic harms and to a globally rising interest in the impact and ethics of algorithms, several, frameworks, tools and guidelines improving algorithmic fairness in the Dutch public sector have already been introduced. This includes an assessment of the impact of algorithms on human rights (Gerards et al., 2022) and a guideline for algorithmic non-discrimination by design (van der Sloot et al., 2021), both commissioned by the Ministry of the Interior and Kingdom Relations[7] and drawn up by scholars at Dutch universities. The latter especially goes into detail about prevention of discrimination according to Dutch law. However, its use is limited to the public sector and because of its non-obligatory nature, the value of this guideline is largely dependent on the intentions and motivation of the organisation using it. To eliminate this dependence on motivation, one could turn to more binding forms of ensuring non-discrimination, such as audits that make both public and private organisations accountable for the potentially discriminatory effects of their algorithms. Hence, this thesis will answer the question: *What role can 'auditing' play in ensuring algorithmic fairness, in terms of non-discrimination?*

This introduction will continue by providing an overview of the scientific disciplines involved in algorithmic fairness research and situate my research within these disciplines. Next, the definition of the terminology, *algorithm, fairness* and *audit,* will be discussed, since all these terms can have different connotations and their meaning fundamentally shapes my research. This introduction ends with an explanation of the approach followed through the rest of this thesis.

[5] Dutch: *Rekenkamer Rotterdam*

[6] More on the difference between direct and indirect discrimination in chapter 2.

[7] Dutch: *Ministerie van Binnenlandse Zaken en Koninkrijksrelaties*

Disciplines involved in algorithmic fairness

The real-world impact of algorithms can be approached from many different directions. Very generally it can be said that the academic fields most relevant to this thesis are computer science, law, social science and philosophy. The attempts to consider the impact of algorithms stemming from computer science itself mostly use the term *algorithmic fairness* (e.g. Kallus et al., 2021; Kleinberg et al., 2018; Pessach & Shmueli, 2023; X. Wang et al., 2022). Important goals throughout computer scientific algorithmic fairness literature are to develop ways (metrics) to quantify and measure the fairness (or amount of undesirable bias) of algorithms and to find (technical) ways to improve this fairness as indicated by these metrics. An important contribution from this perspective is IBMs AI Fairness 360 toolkit, a software package, which contains many state-of-the-art algorithmic fairness metrics and ways to improve it (Bellamy et al., 2018). The quantification branch of the algorithmic fairness approach often takes inspiration from law, since both are concerned with the disparate treatment or disparate impact of algorithms on protected demographic groups. Partly because of this connection, algorithmic fairness has also been studied by law scholars interested in the role algorithmic fairness metrics can play in enforcing non-discrimination legislation (e.g. Hacker, 2018; Hellman, 2020; Nachbar, 2021; Wachter et al., 2020, 2021; Weerts et al., 2023).

Other fields relevant to this thesis, such as Science and Technology Studies (STS) and critical data studies place more emphasize on indirect and fundamental effects of algorithms on society or power structures. The field of STS roughly combines philosophy (mostly ethics of technology) with social sciences (such as anthropology, history, political science and sociology) to investigate the interplay between technology, science and society. A popular framework within science and technology studies is sociotechnical systems theory (e.g. Cooper & Foster, 1971; Fox, 1995; Ropohl, 1999; Whitworth, 2008). According to this theory, the interaction between humans and technology causes both to adapt or be adapted to each other creating sociotechnical systems. A fundamental assumption of sociotechnical systems theory is that, when interacting, humans/social systems and technological systems give rise to emergent behaviour and therefore the holistic analysis of sociotechnical systems adds to the analysis of both the social and technological systems separately (Whitworth, 2008). The sociotechnical view has also often been used to analyse the (social) impact and ethics of algorithms (e.g. Dolata et al., 2022; Draude et al., 2020; Shelby et al., 2023; Shin, 2019) and as a basis for frameworks to audit or assess algorithms (Radiya-Dixit & Neff, 2023; van Bruxvoort & van Keulen, 2021). Furthermore, the cooperations between humans and algorithms that are increasingly prevalent in the workplace can also be analysed as sociotechnical systems.

Critical data studies is another approach which originated from social sciences and can be used to study the impact of algorithms. In contrast to STS, being grounded in critical theory, critical data studies focus on the place of (big) data and the algorithms used to analyse this data in power structures as well as their effect on these structures (Iliadis & Russo, 2016). The wide range of ways in which big data changes or strengthens existing power structures in indirect and intricate ways is mostly beyond the scope of this thesis. Readers interested in this side of algorithmic or big data impact could read Atlas of AI by Kate Crawford (2021).

Since its first edition in 2018, the yearly ACM Conference on Fairness, Accountability, and Transparency (FAccT), has been of key importance in shaping the emerging field of algorithmic fairness. FAccT's approach to algorithmic fairness (and algorithm ethics in general) clearly surpasses a narrow computer science perspective by regarding algorithms as (embedded in) sociotechnical systems (Laufer et al., 2022) and explicitly seeking to bring together "a diverse community of scholars from computer science, law, social sciences and humanities." (ACM FAccT, n.d.).

Big tech companies (a term often used to refer to Alphabet (Google), Amazon, Apple, Meta (Facebook) and Microsoft) are increasingly involved in the discourse surrounding AI ethics as well. Their ethics research groups provided important contributions to the debate on AI ethics and regulatory practices (Crampton, 2022; Croak, 2023; Gebru et al., 2021; M. Mitchell et al., 2019; Pesenti, 2021; Sephus, 2022). However, the good intentions of big tech companies should be viewed with a healthy dose of scepticism, since for these companies the incentive to *appear* moral by having ethical AI research groups, might trump their incentives to *be actually* moral, even if being would counter their goal of maximizing profit. This can result in *ethics washing*, the phenomenon in which companies appear to be moral without making the fundamental changes that would be required for this (Bietti, 2020; Floridi, 2019). This tension between incentives appears to be the reason for the "stochastic parrot controversy", in which Margaret Mitchell and Timnit Gebru, the co-leaders of Google's AI Ethics team, were fired/resigned following their insistence on publishing a paper they co-authored (Bender et al., 2021) which was critical on large language models, a form of machine learning (Dastin & Paresh, 2021; Simonite, 2021)

My own background is in Artificial Intelligence, which is best classified as a computer science background. Given this background, a computer science conception of algorithmic fairness is the most natural starting point for my research. Nevertheless, I will account for the fundamental need for interdisciplinarity in assessing the impact of algorithms by including the sociotechnical systems in which algorithms are embedded in my research. I will do this by informally interviewing practitioners involved in algorithmic fairness assessments about their experiences. Furthermore, I will also connect my research to law by exploring how non-discrimination legislation can inform the assessment of algorithmic fairness. Hence my approach to algorithmic fairness is comparable with the approach that is promoted by FAccT.

What is an algorithm?

Defining the term *algorithm* is notoriously hard, but according to its dictionary definition, it refers to a well-defined, unambiguous method for solving a problem, often executed by a computer (Cambridge Dictionary, n.d.). This definition is mostly sufficient for the purpose of this thesis, but we should add that in context of algorithmic fairness, we are solely concerned with algorithms that solve problems that involve humans, meaning that their output can impact humans in ways that can be considered fair or unfair. Therefore, this thesis will focus on those algorithms producing outcomes that either directly or indirectly lead to decisions with substantial impact on the lives of human individuals. These algorithms are called *decision-making algorithms*, although throughout this thesis they are also simply referred to as algorithms. Examples of decision-making algorithms include algorithms that directly decide whether someone is invited for a job interview and algorithms that calculate a score representing the risk of a bank client participating in money laundering. The latter algorithm can indirectly lead to decisions on whether to subject clients to thorough money laundering investigations or whether to shut down their bank accounts. Decision-making processes that are either directly or indirectly informed by algorithms will be referred to as *algorithmic decision-making* processes. Following Barocas et al. (2023) the persons subject to algorithmic decision-making will be referred to as *decision subjects*.

The decision-making algorithms considered in this thesis are often model-based. In this context, a *model* is a function that can map *input features* to an *outcome value*, also called a *prediction*. Since the algorithms considered decide (possibly indirectly) over humans, the input features will be a selection of those attributes of these humans which are deemed relevant to the problem the algorithm is supposed to solve (provided that these attributes can be translated to data). The outcome is often a prediction about this human or a decision (advice). In case of the job interview invitation algorithm, mentioned above, relevant input features could include highest education received, years of work experience, type of education received, etc. The outcome value could be a binary yes/no decision about whether to invite the applicant or a score representing how suitable the applicant will be for the job or how likely they will be to quit the job soon, etc. Figure 1 shows the main elements of a model-based decision-making algorithm.

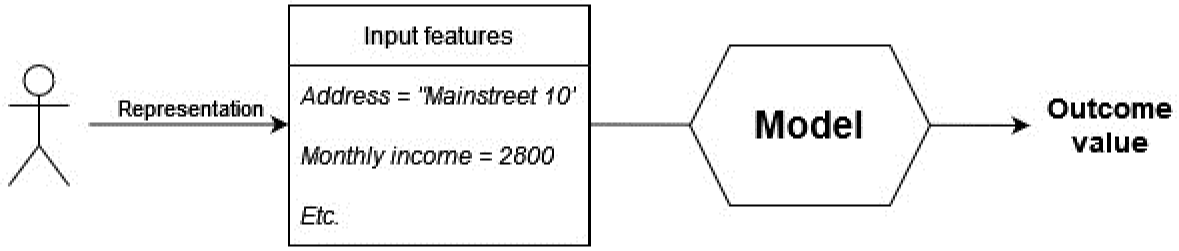


Figure 1: The basis elements of an algorithmic model for decision-making.

Employees of an organisation intending to start using an algorithm could manually design an algorithmic decision-making model by constructing a decision tree or a score-based system in which these employees decide how much each input feature value should contribute to the score that is being kept. (E.g. being highly educated might contribute strongly to a score representing the suitability for a job that requires high intellectual capacity.) However, an increasingly popular alternative is to use *machine learning* (ML). When using ML, models are fitted on large data sets that consists of many instances of all relevant input features (of the same individual) and the corresponding "true" or "right" output value (the *target value*)[8]. During this training process, the ML model is adapted in such a way that it will be more likely to produce an accurate output value when given the corresponding input features. The model learns to detect patterns in the input data that statistically correlate with certain output values.[9] In public discourse and some literature, the term *Artificial Intelligence* (AI) is used interchangeably with ML, although AI is more of an umbrella term that includes other attempts of making machines behave or operate humanly or rationally (Russell & Norvig, 2010, pp. 1–2). The term *deep learning* refers to a specific type of ML models (multilayer artificial neural networks) that excel in approximating complex input-output relations. Deep learning models are notorious for their so-called opaqueness. It is practically impossible to explain how a deep learning model derived its output from a given input and how it works in general, which is why deep learning models are often referred to as *black boxes* (Lecun et al., 2015; von Eschenbach, 2021). This black box nature of some algorithms can make it hard to judge whether they are discriminatory, since this judgement might require knowing whether a person was treated less favourably *because of* them being black or female or queer or old, etc. More on different definitions of discrimination in the following chapters.

What is fairness?

At the start of this introduction, we already related algorithmic fairness to non-discrimination. In this section, we go more in depth on this choice and the many meanings fairness could have.

[8] For many decision problems a "true" or "right" outcome value does not exist or cannot easily be found. (E.g. what is the right option when deciding whether to invite an applicant for a job interview?) In these cases, we let the algorithm predict a value that is supposed to be indicative of the "rightness" of a decision. More on that in the section Shortcomings of algorithmic fairness.

[9] Readers who got lost in this short, information-dense explanation of ML, might be helped by watching explanation videos that can be found by searching for "machine learning basics" on video sites such as YouTube. Readers who want to go more into dept into different types of ML and its mathematical details can consult books such as (James et al., 2021; Theodoridis & Koutroumbas, 2006).

Fairness as essentially contested concept

The placement of the datacentres needed to run (and, in case of ML, train) the algorithms, the environmental effects of these datacentres, the labour conditions under which data essential to the algorithm was annotated, the manipulative power that the design of the algorithm might grant to its deployer and many more elements in the construction, use and impact of algorithms could all lead to some form of harm or unfairness. A taxonomy of all harm that could be done using algorithms was proposed by Shelby et al. (2023). The reason why the range of ways in which an algorithm could be considered unfair is not only that the development, output and maintenance of algorithms can impact such a wide range of people: the inherent vagueness of the term fairness is what makes it possible to call a wide range of practices unfair.

As Nachbar (2021) argues, fairness is an *essentially contested concept*, a term originally introduced by Gallie (1955) to denote abstract and evaluative notions, such as work of art, democracy or Christian doctrine, which do not have an undisputed, universal definition. A statement that does not contain any essentially contested concepts, such as “Neil Armstrong set foot on the moon” can only be disagreed on by two persons who have different believes about the factual state of world. In contrast, a statement such as “Obamacare is fair” can be disagreed upon even by two people who agree on the facts about what Obamacare entails but hold a different notion of the essentially contested concept of fairness.

It is certainly possible give a specific definition of fairness or algorithmic fairness more specifically. For example, we could state that an algorithm is fair if it ultimately promotes the happiness of all parties it impacts (Bentham’s utility principle). In contrast to fairness as a *concept*, such a specific definition of fairness can be called a *conception* of fairness. Dworkin (1972) argues that due to fairness being an essentially contested concept, those responsible for setting “standards of fairness” are faced with two options: they could either appeal to the ideal of fairness by simply demanding that people (or in our case, organisations using algorithms) act fairly, without further clarification, or they could define a specific conception of fairness, which everyone should adhere to. Within the current landscape of data politics, different parties use different conceptions of fairness while others seem to appeal to a vaguer ideal of fairness. This lack of consensus increases the risk of ethics washing (Bietti, 2020; Floridi, 2019).

The consequence of fairness being an essentially contested concept is that each attempt to test for fairness will automatically specify the concept to a certain extent. Imagine a factory owner, called Solomon, who wants to assess whether his factory treats its employees fairly, without any specification of the meaning of fairness. Solomon might start by inspecting the wages employees receive, their workload, the policy for illness, etc. Whatever he decides to inspect and leave uninspected can be seen as an interpretation of the concept of fairness. Apparently, fairness is respectively concerned with the domain of money, labour or treatment of the ill. And when one or several domains are chosen for inspection, more choices will arise. If Solomon decides to inspect the wages, will he inspect whether wages are proportional to the work that was performed, whether wages are equal across workers who perform similar work, etc.? And if Solomon choses to inspect whether wages are proportional to the work that was performed, how does he measure this proportionality?

After Solomon has made all these decisions, we might still not have a precise conception of fairness, but at least we know that (according to him) a factory that -for example- pays all employees at least a minimum wage, pays workers with at least five year working experience at least 1.5 times minimum wage, provides paid leave in case of illness and pregnancy and has working days of eight hours is fair. With this judgement, Solomon has at least ruled out some of the perhaps infinite possible conceptions of fairness, such as a conception that says that all people with the same job should get equally paid regardless of differences in experience. He might not have been aware of his contribution to promoting and devaluating certain conceptions of fairness, but in judging whether anything is fair such a promotion and devaluation cannot be avoided. In other words: assessing whether an essentially contested concept (e.g. fairness) applies to a certain phenomenon (e.g. a factory) is inherently normative, if there is no agreed upon conception of this concept against which the phenomenon can be objectively measured. The choices faced in such an assessment of an essentially contested concept that promote certain conceptions of this concept while devaluating others (thereby shaping the interpretation of the concept) will be referred to as *normative assessment choices*. These are related to what Hildebrandt (2020) refers to as design choices, the main difference being that she focusses on *algorithmic fairness by design,* and we focus on algorithmic fairness assessments and audits[10]. Hildebrandt stresses that these choices have moral implications.

Conceptions of algorithmic fairness

As stated above, within computer science literature algorithmic fairness is often defined in relation to the disparate impact or treatment of an algorithm. Here, the literature ultimately measures fairness in the outcome values of the algorithm. This focus is suitable if we are interested in matters of non-discrimination or "fair" treatment but excludes more indirect and abstract harms that could result from the use of algorithms from the domain of algorithmic fairness. Think of the examples such as the environmental costs of training ML (if used) or the labour conditions of those preparing data for the algorithm. It is clearly beyond my expertise and arguably impossible to account for all ways in which the development, deployment and use of algorithms could promote or harm fairness. On the other hand, purely limiting my discussion to a prevalent conception of algorithmic fairness in computer science, seems rather senseless and is perhaps impossible as well, since in choosing a conception of algorithmic fairness from computer science literature normative assessment choices will inevitably be made.

Balancing the need to consider algorithmic fairness from perspectives beyond computer science and the need for a well demarcated research topic, which can be approached well from a computer science background, I decided to depart from computer science in defining algorithmic fairness and search for interdisciplinary input from that perspective. Most conceptions of fairness in computer science literature are phrased in terms of bias or prejudice. According to Mehrabi et al. (2021, p. 155:2), for example, fairness is "the absence of any prejudice or favouritism toward an individual or group based on their inherent or acquired characteristics." However, to require a fair algorithm to be completely free of bias is not sensible. As Hellström et al. (2020) rightfully note, the very purpose of decision-making algorithms is to select some persons and not select others. Unless this selection would be completely at random (which would render the algorithm rather useless), it will always need to be based on input features and in case of decision-making about persons or groups these input features will often be characteristics of these persons or groups. Hence, the issue at stake with (algorithmic) fairness is not whether an algorithm is free of bias, but whether it is free of *unwanted bias*. For example, if an algorithm detecting potential fraudulent clients of a certain bank is biased against clients who transfer large amounts of money to bank accounts known to be involved in money laundering (in the sense that the algorithm predicts these clients to be more likely to be fraudulent) this is commonly accepted as fair, but if this algorithm would be biased against people with a certain sexual orientation this would commonly be considered unfair.

If improving algorithmic fairness is taken to mean reducing unwanted bias, the obvious question is how we can determine whether bias is unwanted. Unfortunately, this can poorly be formalised since being unwanted is purely subjective. This is why "unwantedness" has often been linked to (legal)

[10] No matter whether we design an algorithm in such a way that it satisfies a certain conception of fairness or whether we assess it so that we only accept algorithms that satisfy a certain conception of fairness, we will face similar or equal choices.

discrimination: if bias is discriminatory, it is considered unwanted (X. Wang et al., 2022). Formalizing whether bias is discriminatory seems more promising. Although the concept of discrimination might have different conceptions as well[11], national and international non-discrimination legislation force a more or less fixed conception of fairness upon society that could be used to decide when a practice should be considered discriminatory and when not. This thesis will aim to find a way of assessing algorithmic fairness, conceptualised as non-discrimination, since grounding our conception of fairness in law seems the only way to avoid either forcing an arbitrary conception of fairness upon algorithm deployers or leaving it up to them individually how to interpret the essentially contested concept of fairness. Furthermore, in turning to law, my approach will not be isolated in computer science, as I will make a connection to society, mediated by law.

In short, the working definition of algorithmic fairness for the purpose of this thesis, will be the absence of discrimination by an algorithm. How the absence of discrimination (or non-discrimination) can be conceptualised will be discussed in chapter 2. It should be noted that with aligning our current conception of fairness with non-discrimination, we limit our scope to only a small fraction of the harmful impact that algorithms could have on persons, groups or the environment. In fact, this choice of focus could be considered a normative assessment choice. In conceptualising algorithmic fairness as non-discrimination, I do not mean to suggest that this is truly all there is to this topic. Research into other aspects of algorithmic fairness and how they can be assessed or potentially audited is welcome but is better conducted by scholars with more knowledge of social sciences and sociopolitical systems.

What is an algorithmic fairness audit?

Currently, there appears to be a rising interest in finding ways to regulate the use of (decision-making) algorithms. An important tool that is often suggested to serve this aim is the *algorithm audit* or sometimes specifically an AI or ML audit (e.g. Brown et al., 2021; Raji et al., 2020; Sandvig et al., 2014; Spielkamp, 2023; The Supreme Audit Institution of Finland et al., 2023). The term auditing originates from the context of (financial) auditing. According to its ISO definition, an audit is a “systematic, independent and documented process for obtaining objective evidence and evaluating it objectively to determine the extent to which the audit criteria are fulfilled.”, where audit criteria are a set of requirements, that might be legally prescribed, generally implied as common practice or stated in advance. (International Organization for Standardization, 2018). It is essential that the fulfilment of audit criteria only depends on objective evidence. Requiring candy manufactures to only produce tasteful candies would be an unfit audit criterium since tastefulness is subjective. Requiring candy manufacturers to have all their types of candies tested and approved by an independent testing committee, however, is better suited as audit criterium since compliance can be objectively measured. In other words, non-objective, context dependent or ambiguous requirement such as tastefulness should be operationalised by finding objective audit criteria that can serve as substitute. Although below it will become apparent that the strict ISO definition of audits is often watered down when it comes to algorithm audits, it is still the definition that will be used for the purposes of this thesis.

It is common to distinguish between internal and external audits as well as between first party, second party and third-party audits. The terms *internal audit* or *first party audit* often refer to audits conducted by the audited organisation (*auditee*) itself or by an external auditor commissioned by the auditee. *Second party audits* are performed by or on behalf of parties that have a contract with the auditee, such as customers who want to make sure the parties, they cooperate with meet certain requirements. Confusingly, sometimes the term second party audit is also used to refer to an audit by an external auditor commissioned by the auditee, which we classified as a first party audit. *Third party audits* are

[11] See the chapter 1 for different forms or possible meanings of discrimination.

performed by auditors who are completely independent from the auditee, such as accountancy or consultancy firms or government bodies. Together, third- and second-party auditing is commonly referred to as *external auditing* (International Organization for Standardization, 2018).

Although auditing is often associated with the financial domain, in which the frameworks and standards backing up a trustworthy and healthy auditing system are arguably most mature, the definition above is agnostic towards the domain and underlying purpose of audits if there are audit criteria which can be tested against objective evidence. Indeed, over centuries, the practice of auditing has shifted from being solely concerned with fraud detection to keeping in check the overall organisational structures of companies. (Md Ali & Teck-Heang, 2008). As companies started to rely increasingly on IT systems, monitoring the robustness, stability, safety and quality of these systems naturally became a branch within auditing as well (Barta, 2018). The current rise of algorithm auditing could be portrayed as a natural continuation of this trend, given that today an increasing number of companies relies on algorithmic applications.

Depending on the audit criteria chosen, audits of algorithms could take vastly different forms. Given our current focus on assessing algorithmic fairness, these criteria should be concerned with the impact of algorithms. These audits are often called *ethical* algorithm (or AI or ML) audits (Brown et al., 2021; Mökander & Floridi, 2021; Rai, 2021; Zinda, 2022). The part of ethical algorithm auditing my thesis will focus on is algorithmic fairness auditing. Audits focussed on algorithmic fairness often report the performance of an algorithm on different fairness metrics. (See chapter 1.)

Auditing algorithmic fairness in practice

Some audits of algorithms partially covering or dedicated to fairness have already been executed. Most of them had a technical focus, meaning that performance on one or several fairness metrics was reported. (See chapter 1 for more on fairness metrics.) Other audits have a social or organisational focus, for example assessing how the organisation involves stakeholders to assess social impact beyond technical fairness metrics or if the internal policy a company must ensure fairness suffices.

Examples of self-acclaimed algorithmic fairness audits from academia include an independent technical audit of commercially available facial gender recognition technology (Buolamwini & Gebru, 2018) and a sociotechnical audit of facial recognition technology used by the police (Radiya-Dixit & Neff, 2023). The former is not an audit according to its strict ISO definition, since it did not depart from a pre-established set of audit criteria.

Sometimes audits are also conducted by both big accountancy firms traditionally focussed on financial audits and new consultancy firms specifically focused on algorithms and their impact. These audits are often, based on self-produced auditing frameworks. Although these audits are sometimes referred to as third-party, since they are executed by an independent party, they are arguably first-party audits because most often they are initiated/requested by the auditee. Often the reports resulting from these audits are not publicly disclosed, although examples of published audit reports include a technical and organisational largely fairness focused audit of an algorithm used for selecting job applicants (ORCAA, 2020) and an organisational audit of Facebooks commitment to civil rights which included assessing Facebooks policy for ensuring the fairness of their algorithms (*Facebook's Civil Rights Audit-Final Report*, 2020). However, both audits do not satisfy the ISO requirement of using a pre-established set of audit criteria.

Existing governmental oversight institutions can also execute fairness audits. The Netherlands court of audit[12] developed its own framework for AI auditing which includes fairness in its scope. This framework was applied to audit nine algorithms used by the Dutch government. This framework contains a set of criteria and therefore it is an important step in enabling algorithm audits that fulfil the strict ISO requirements for audits. Unfortunately, however, the exact meaning of fairness within this framework is not clearly defined, making it impossible to assess the fairness related requirements in this framework objectively (Algemene Rekenkamer, 2020, 2022). Therefore, these requirements are unfit as proper audit criteria.

Lastly, several journalists have investigated algorithmic fairness in a way that resembles the methodology of technical audits as they are currently executed (although not using pre-established criteria either). Examples included investigations in the fairness of an algorithm used by the municipality of Rotterdam to detect cases of people unjustly receiving welfare (Geiger et al., 2023) and an algorithm predicting the chance of recidivism for offenders used by US lawyers to inform their verdict (Angwin et al., 2016; Larson et al., 2016).

The risk of audit washing

Perhaps unfortunately, regulation generally does not mandate specific (algorithmic) fairness requirements that can be straightforwardly translated to audit criteria. A notable exception is upcoming New York legislation that mandates audits using a specific fairness metric to avoid unwanted bias against protected groups in algorithms used to assist employment processes (Cumbo et al., 2021). In EU context the only legal requirements that could serve as grounds for algorithmic fairness audits are to be found in non-discrimination legislation, which prohibits discrimination regardless of it being caused by men or machine. (More on this in chapter 2.) However, currently it appears to be up to the auditor to decide how to test or audit for fairness or non-discrimination. This situation introduces a great risk of *audit washing*. Similar to green washing and ethics washing, audit washing is the process of acquiring audit verifications from audits which are defined or executed poorly. These verifications have little to no meaning and could potentially distract from the actual moral harm done by the auditee (E. P. Goodman & Trehu, 2022).

In response to the risk of audit washing, some argue for more precise standards and regulations backing up audits, answering the what, who, why and how of auditing and/or mandating audits (Costanza-Chock et al., 2022; E. P. Goodman & Trehu, 2022; Lucaj et al., 2023; Mökander et al., 2022). The ForHumanity institute aims to provide a universal framework for AI auditing which includes fairness. At the same time, they advocate to restrict the definition of AI audits such that critical investigations only qualify as audit if they are executed by certified and independent practitioners who objectively assess conformity to binary rules, have to face consequences for false audit certificates themselves and do not provide feedback to the auditee in any form other than a final public report (Carrier, 2021; Carrier & Brown, 2021). In short, they want AI audits to be closer to the ISO definition of audits. The AI NOW institute however, fears that developing coherent auditing frameworks is so challenging (especially for complex and powerful big tech platforms) that audits are more likely to "devolve into a superficial 'checkbox' exercise." (Kak & Myers West, 2023, p. 37) Because of this, AI NOW wants to move beyond AI auditing all together to focus on more fundamental ways of limiting big tech power (e.g. antitrust laws) and ensuring sound AI use.

We appear to have discovered a tension between those who think algorithmic fairness can and should be formalised and tested and those who think the conception of fairness the former group aims to test for is too limited and diverts attention from the real problems of algorithms and AI. This tension

[12] Dutch: *Algemene Rekenkamer*

resembles the tension described above in this thesis, where I had to balance the need to consider algorithmic fairness in all its complexity and from perspectives beyond computer science (the elements of fairness AI NOW points to) and the need for a well demarcated research topic approachable from a computer science perspective (which would be more in line with attempts to formalise fairness using metrics and audits). Again, my answer is that auditing algorithms for a limited conception of fairness as non-discrimination certainly is not all there is to ensuring that the use of algorithms will be fair and ethical, but that it still could have function in this large and complex task. This does mean that the limited scope of algorithmic fairness audits or assessments should be always remembered. What the role of auditing could be exactly, is the subject of this thesis.

Approach

In this introduction we showed that the use of algorithms in decision-making processes can lead to discrimination. Computer science literature offers metrics for measuring certain kinds of (discriminatory) bias in algorithms. These metrics are sometimes used for self-proclaimed algorithmic fairness audits, although these audits often lack a set of pre-established, objective audit criteria to depart from. If the aim of algorithmic fairness audits is to establish with certainty that an algorithm has not discriminated, this set of audit criteria should guarantee compliance with non-discrimination legislation. Perhaps unfortunately, non-discrimination itself does not appear to be an objective criterium that can be assessed objectively. However, it might still be possible to define a set of objective audit criteria, which together suffice to ensure non-discrimination, when all these criteria are fulfilled. Such a set of criteria, together with strict rules about how and by whom the audit should be performed and how it should be documented, is what we will call an *audit framework*.

Even if constructing an audit framework which guarantees non-discrimination turns out to be too ambitious, auditing might still be a valuable instrument in reducing the risk of algorithmic discrimination. For example, even though it might be impossible to define a set of audit criteria, which form a sufficient requirement for fairness when satisfied, it might still be possible to define audit criteria which are necessary requirements for non-discrimination. This means that if an audit for such criteria is failed, the algorithm certainly discriminates and if it is passed, the algorithm meets important requirements for non-discrimination, but might still discriminate in a way that the audit cannot detect. Alternatively, it might be possible to audit whether a certain test for algorithmic fairness was executed properly, whereas it might not be possible to audit whether the right conclusions were drawn from the results of this test, since the latter question might depend on subjective interpretation. Because there are different ways in which audits could be used in ensuring non-discrimination, the question of *how* (or in what role) auditing can be used for this aim, seems more relevant than the question *if* it can be used, which is why we phrased our research question as: What role can *auditing* play in ensuring algorithmic fairness, in terms of non-discrimination?

In investigating our research question, we take the following steps. Firstly, we need to familiarise ourselves with the technical methods for assessing algorithmic fairness to identify the normative assessment choices involved in this. For this purpose, the first chapter of this thesis will investigate algorithmic fairness from a computer science perspective. The fact that decisions (normative assessment choices) will inadvertently be faced when assessing algorithmic fairness, does not necessarily rule out the possibility that fairness assessments could be completely captured in an audit framework. After all, if the law clearly prescribes the preferred option for each question faced during such assessment, constructing such a framework is still possible. This is why the second chapter of this thesis will look for an answer to these questions in Dutch non-discrimination legislation. Finally, the third chapter investigates the human part of the sociotechnical systems in which algorithms are embedded by showing conversations with practitioners involved in assessing algorithmic fairness. This chapter provides examples of how normative assessment choices were made relying on non-discrimination legislation and a wide range of different sources.

Chapter 1: Algorithmic Fairness

To find out what role auditing can play in assessing algorithmic fairness we need to investigate the possibility of capturing such assessments in audit criteria. Since audit criteria need to be binary and objective, there is much value in quantifying algorithmic fairness. After all, if a suitable fairness metric exists, all an audit criterium must demand is that an algorithm scores above or below a certain threshold on this metric. Fortunately, computer scientists generally love quantifying problems (since quantification is also needed for automating optimisation) and hence a large body of computer scientific algorithmic fairness literature is dedicated to finding suitable algorithmic fairness metrics. Therefore, we need to find those fairness metrics that are most suitable for preventing algorithmic discrimination.

This chapter will introduce the computer scientific field of algorithmic fairness and then it will summarise the most important methods for measuring fairness in this field and discuss the underlying conceptions of fairness. It first discusses how algorithmic fairness can be analysed by looking at an algorithm's input and then how it can be analysed by looking at the output. Next, this chapter will discuss the implications of most fairness assessment methods not directly showing the real-world impact of algorithms and the limitations this causes. The chapter will end with a summary of how algorithmic fairness theory can be of use in showing the demographic impact of algorithms including a list of the most common metrics available, their uses and shortcomings. The normative assessment choices that should necessarily be made in assessing fairness will also be identified and summarised.

Introduction to algorithmic fairness

As discussed in the introduction of this thesis, decision-making algorithms are parts of sociotechnical systems, and their use will have social impact. Here, the term social impact refers to a great range of impacts the algorithm can have on a great range of social or demographic groups (Freudenburg, 2003). When this impact disadvantages a person or group, it can also be called a harm. The harms that are most directly related to discrimination are called *allocative harms* by Shelby et al. (2023). Allocative harms consist of loss of opportunity and economic loss. Table 1 provides examples of both types of harms caused by hypothetical algorithms. Notice that loss of opportunity or economic loss do not always point to unfairness. If a driver loses their licence due to an enormous exceedance of a speed limit, this results in opportunity loss for the driver, since they are no longer allowed to drive, limiting their freedom of movement. However, most people will not consider this unfair, if the loss of the driver's license was justified and all drivers are equally likely to lose it, if they exceed speed limits by enormous amounts. Hence, the issue at hand appears to be less about suffering harm or being benefitted and more about whether it was justified or fair that a decision subject got to experience this harm or (lack of) benefit. Unfortunately, this means that we end up where we started, facing the question what it means for an algorithm -or in this case an allocation of benefit or harm- to be fair.

Individual fairness and group fairness

Within computer science, there are two major approaches to answering this question, individual fairness and group fairness. The *individual fairness* approach poses that an algorithm (or more generally: a decision-making process) is fair when "similar" individuals receive "similar" decision outcomes. Similarity between individuals can be expressed by a mathematical function that compares the attributes of individuals and adds to a difference score when attributes do not match. The lower the final difference score, the higher the similarity between the individuals (Binns, 2020; Mukherjee et al., 2020). This approach would require deciding how much each difference between attribute values should contribute to the difference score and what the relationship between the difference score for two individuals and the maximally accepted difference between their outcome values should be

(Dwork et al., 2012). These decisions are far from straightforward and will always lead to a somewhat arbitrary demarcation between fair and unfair decision-making processes.

The second approach to algorithmic fairness is *group fairness*. Instead of determining whether an individual decision outcome was deserved, this approach requires that people with different demographic characteristics are not treaded or impacted considerably different by an algorithm. In other words, group fairness is satisfied when a decision-making process does not discriminate (where the meaning of discrimination is open to debate). This is why we will focus on this conception of fairness in this thesis. Discrimination is often defined in relation to *protected groups*. Protected groups are groups of people that share a *protected attribute* (also called sensitive attribute), which is a personal trait that has historically been used as ground for discrimination (such as sex, sexual orientation, religion or ethnicity) and is now legally protected from serving as a ground of discrimination.[13] (e.g. Coston et al., 2019; Ravfogel et al., 2020; Romanov et al., 2019; Yu et al., 2021). Based on social, economic and historical context, we could divide protected groups into historically *privileged* group(s) and *unprivileged* groups. In context of a patriarchal society, the protected attribute sex, for example, would have males as privileged groups and females (and possibly intersex people) as unprivileged groups. The goal of promoting group fairness could be framed in two ways: the goal could be to prevent discrimination based on protected attributes by the algorithm altogether or the (more ambitious) goal could be to contribute to a society in which the position of unprivileged group members is more equal to the position of privileged group members, possibly compensating for existing social inequalities. Assuming the latter goal, treating people of privileged and unprivileged groups differently in a way that benefits people of unprivileged groups (often called positive discrimination) is permissible.

Groups can either be discriminated directly or indirectly. We speak of *direct discrimination* when the decision to treat someone differently is directly based on a protected attribute this person possesses. *Indirect discrimination*, on the other hand, occurs when seemingly neutral rules or actions not directly based on protected attributes, still end up discriminating against certain protected groups (e.g. Campbell & Smith, 2023; Hajian & Domingo-Ferrer, 2013; Maliszewska-Nienartowicz, n.d.; Zhang et al., 2017).

Favourable versus unfavourable outcomes

To better understand the meaning of group fairness in algorithmic context, it is important to get a better understanding of the specific algorithms we are concerned with. Although this is common knowledge in computer science and ML, I will still briefly outline this terminology, as this thesis is addressing a interdisciplinary audience. Most algorithms used in decision-making are *classification* algorithms. These are algorithms that return one of a fixed set of possible classes as output value. *Binary classification* algorithms can only return one of two possible classes. E.g. a binary classification algorithm used for job applications could classify CVs as either suitable or not suitable for the job. In binary classification, we often distinguish between the class that we primarily want to select (in the example: suitable applicants) and the class of all instances that do not belong to this former class (in the example: unsuitable applicants). The former class is often called the *positive class* and the latter the *negative class*.[14] *Multiclass classification* algorithms use more than two classes. E.g. if our job application algorithm was a multiclass classification algorithm, it could, for example, classify CVs as very unsuitable, probably unsuitable, probably suitable or very suitable for a job. Embedded in a decision-

[13] More on the role of protected attributes in non-discrimination legislation in chapter 2.

[14] In a certain sense, distinguishing the positive class from the negative class is a matter of interpretation and somewhat arbitrary. If we would say that the primary aim of the algorithm in our example is to detect CVs of *un*suitable candidates so that they can be deleted without wasting effort on them, the positive class and negative class would switch places.

making process, very suitable and very unsuitable CVs might respectively be accepted for a job interview and rejected without human intervention, whereas CVs in the two intermediary categories will be evaluated by humans.

In addition to classification algorithms, *regression* algorithms could also be used in decision-making processes. In contrast to classification algorithms, regression algorithms return a numerical value, often within a fixed range. This value can often be interpreted as the estimated probability that a data instance will belong to a certain class. E.g. a regression algorithm might be used in fraud detection to estimate the risk that a person is fraudulent. The probability or risk value estimated by the algorithm can be used as input by a human decision maker in deciding whether the decision subject should be subjected to a thorough fraud investigation. Alternatively, a threshold could be set, such that all decision subjects with a risk value that exceeds this threshold should automatically be subjected to further investigation or the X decision subjects with the highest risk scores will be selected for this. These last two options effectively reduce the regression algorithm to a binary classifier, since in the end all decision subjects will either be investigated (the positive class) or not (the negative class).

The methods discussed in this chapter only apply to binary classification. That is to say, they rely on a clear, binary distinction between outcomes that are *favourable* (e.g. being invited to a job interview or *not* having your bank account being shut down) and *unfavourable* (e.g. having your bank account being shut down or *not* being invited to a job interview) for the decision subject (Lepri et al., 2018; R. Wang et al., 2020). The terms *favourable group* and *unfavourable group* will be used to refer to the groups of people for which the algorithmic decision process respectively resulted in a favourable or unfavourable decision, irrespective of whether this was the "right" decision. The favourable group should not be confused with the positive class (and the unfavourable group should not be confused with the negative class). E.g., in case of a fraud detection algorithm all decision subjects suspected of fraud will be assigned to the positive class, even though these individuals clearly would not consider this suspicion to be positive (if we take positive to mean favourable). For each example, Table 1 shows whether an unfavourable outcome corresponds to a negative classification or a positive classification.

Harm type	**Example of algorithm causing this type of harm**	**Unfavourable outcome in example**	**Unfavourable outcome corresponds to**
Opportunity loss	An algorithm used to admit students to a prestigious university programme.	An aspiring student is excluded from the programme.	Negative classification
	An algorithm used to flag welfare recipients who are suspected of potential fraud.	A welfare recipient is rejected further welfare.	Positive classification
Economic loss	An algorithm used in offering personalised discounts to web shop costumers.	A costumer does not receive a discount.	Negative classification
	An algorithm used to detect offensive content on a video sharing site and exclude it from making revenue from advertisements.	A content creator does not receive income from people watching their video.	Positive classification

Table 1: Algorithmic harms relevant to fairness. *This is a selection of the harms found in the taxonomy by Shelby et al. (2023).*

A focus on decision-making processes which result in a clear, binary distinction between favourable and unfavourable groups, is less limiting for the applicability of our analysis than it might appear at the surface. Even if the algorithm used in a decision-making process is not a binary classifier itself, the decision-making process (a sociotechnical system) might result in a binary classification nevertheless. For example, non-binary algorithmic output (such as a risk score or a risk indication class) might be one of the factors considered by a human decision maker when making an ultimate, binary decision.

Furthermore, even when a decision is not binary, it might still be possible to identify one or several clearly favourable and unfavourable decision outcomes. E.g. imagine a decision process serving to detect welfare fraud has three possible decision outcomes: (1) the decision subject is left alone and does not experience any negative consequences, (2) the subject is invited for a one-to-one meeting with a government official with the aim to clarify the situation or (3) the subject immediately has to pay a fine and the welfare is stopped until further investigation has ended. In this situation it might not be entirely clear whether option 2 should be considered favourable or unfavourable for the decision subject. However, it appears very clear that option 1 is favourable and option 3 is unfavourable. The methods discussed in this chapter will work perfectly fine if the favourable group and unfavourable groups respectively consist of all decision subjects to whom option 1 applied and to whom option 3 applied, ignoring all people to whom option 2 applied. The only requirement here is that the favourable group and unfavourable group are sufficiently large for most methods to be reliable. Additionally, a division between favourable and unfavourable groups aligns well with the legal non-discrimination framework as described in chapter 2.[15]

Causes of algorithmic discrimination

A large portion of the computer science literature on algorithmic fairness is dedicated to measuring unwanted or discriminatory bias. However, if we could prevent algorithmic discrimination in the first place with certainty, measuring it is not necessary. Algorithmic discrimination can often be traced back to societal or institutional patterns of discrimination. In case of ML, the algorithm might learn to replicate prejudices that are captured in the data it is trained on. The phenomenon of ML model output being of low quality because the data the model was trained on is of bad quality, is often summarised as *garbage in, garbage out* (E.g. Canbek, 2022; Hyde et al., 2023; Stuart Geiger et al., 2020). In context of algorithmic fairness, this saying has been adapted to *bias in, bias out* (Mayson, 2019), which refers to the fact that if the data used to train a ML model contains (discriminatory) biases, the model is likely to learn to replicate or possibly even magnify these biases. An example of this principle in action is the algorithm used in Amazon's hiring practices, which learnt to discriminate against women because the target labels in the data it was trained on originated from human, biased selection committees (Dastin, 2018). In case of humanly designed algorithms, humans might directly base the design of the algorithm on (possibly subconscious) prejudices. E.g. the algorithm used by the Dutch executive government body for education to detect student grant fraud was designed to assign higher risk scores to student living with family members outside of their original households (e.g. uncles, aunts, cousins or grandparents). Since this is especially common in some cultures of Dutch citizens with migration backgrounds (regardless of potential fraud), this caused (arguably discriminatory) bias against decision subjects with a migration background (Belleman et al., 2023).

Especially in case of (deep) ML it might be very hard or impossible to control the rules that are captured in an algorithmic decision-making model and prevent any of them from being discriminatory. Furthermore, bias in training data labelling might sometimes be hard to detect. However, it is relatively easy to control the input features of algorithms. Hence, we might try to prevent algorithmic discrimination by removing all input features that might result in it. A logical starting point would be to make sure no protected attributes are used as input features. This is arguably sufficient to prevent *direct* discrimination, since in algorithmic context, direct discrimination appears to be only possible if a protected attribute is included in the input features of the algorithm, so that the algorithm can use it to derive its decision outcome. Hence if the input features of an algorithm do not contain any protected

[15] Readers who are interested nevertheless in methods for testing algorithmic fairness beyond binary classification, can find them elsewhere (Blakeney et al., 2022; Steinberg et al., 2020; Verma & Rubin, 2018).

attribute, this algorithm satisfies the fairness definition known as *fairness through unawareness* (e.g. Cornacchia et al., 2023; Kusner et al., 2017; Verma & Rubin, 2018).

Unfortunately, however, this method does not prevent indirect discrimination, which, in some cases, might be just as harmful as direct discrimination (Kearns & Roth, 2019, Chapter 2). For example, an algorithm that systematically more frequently produces unfavourable outcomes for people who possess the attribute "having been to Mekka at least once", does not directly discriminate against Muslims, since it does not use religion as an input feature. However, the causal link between being Muslim and possessing this attribute is so strong that it would be intuitive to say this algorithm is discriminatory. In this case discrimination is not directly based on protected attributes, but on proxies of these attributes (visiting Mekka). This is why this form of indirect discrimination is also called *discrimination by proxy*. (e.g. Alexander, 1992; Alexander & Cole, 1997; Hajian & Domingo-Ferrer, 2013; Johnson & Martinez, 1999; Prince & Schwarcz, 2019). Common proxies for protected attributes include ZIP code for race or nationality (since many cities are racially segregated) or working hours for sex or gender (because women more often work parttime).

Whether attributes are proxies for a certain protected attribute is highly context depended. (E.g. in a city with little racial segregation, ZIP code would not be a proxy for race.) Furthermore, according to a strict definition of discrimination by proxy, an attribute only serves as proxy if the predictive value of the feature is largely derived from its relation to the protected attribute, instead of a direct relation between the input feature and outcome not involving the protected attribute (Prince & Schwarcz, 2019). Figure 2 shows when an input feature can cause discrimination by proxy (such as the example of having visited Mekka as proxy for being Muslim) and when it does not. The rationale behind this is that in diagram A the input feature is merely used as a stand-in (or proxy) for the protected attribute, enabling the replication of an existing discriminatory link between the protected attribute and the decision outcome. (In the hypothetical example this link would be discrimination of Muslims, lowering their chances of admission to a university.) In diagram B, however, the algorithm directly uses the predictive power of the input feature and not the fact that it could also serve as proxy. Yet, even in situations that can be modelled as diagram B, the result will be that people of different protected groups receive different outcomes, which in some contexts might still be undesirable. Whether cases of indirect discrimination should be strictly limited to cases of discrimination by proxy, or we should speak of indirect discrimination in cases that can be modelled by figure 2.B, without speaking of discrimination by proxy, depends on our definition of indirect discrimination.[16] This will be examined in more detail in chapter 2.

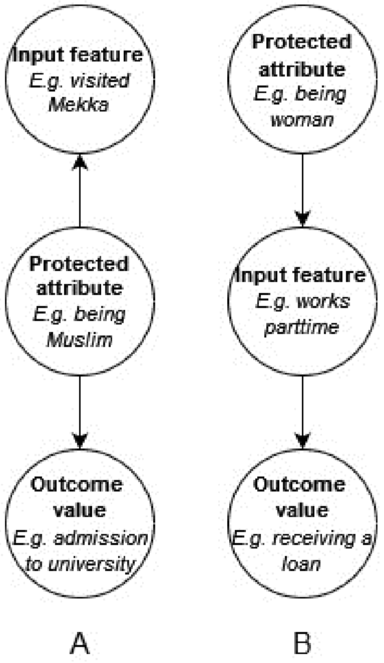


Figure 2: Discrimination by proxy. *Two diagrams showing possible relationships between an (unobserved) protected attribute, an input feature and an outcome value. The arrows indicate causal pathways that explain the influence of certain personal attributes on others, within a given society. (E.g. in societies with institutional discrimination of Muslims, being Muslim might (in many different and complex ways) casually influence the likelihood of being admitted to a university.) Only for diagram A the use of the input feature satisfies a strict definition of discrimination by proxy.*

No matter how we define indirect discrimination, it is apparent that *fairness through unawareness* does not provide any

[16] Causal reasoning fairness metrics explicitly draw on causal graphs (such as the one shown in figure 2) and their implications for (un)fairness (Verma & Rubin, 2018). However, they are beyond the scope of this thesis.

certainty of preventing it, since the input features of the algorithm might still contain proxies of protected attributes or input features correlated to protected attributes that are not strictly proxies. A response to this problem might be to attempt to remove all sources of indirect discrimination from the input features as well. For this purpose, we would first need a way of identifying these sources of indirect discrimination.

Firstly, this could be done quantitatively by calculating the correlation between each input feature and all protected attributes considered. The method of calculating these correlations and removing the input features most correlated with a protected attribute is called *suppression* (Kamiran & Calders, 2012). However, this can be done only if data on these protected attributes was collected and stored. A calculation of the corelation between input features and protected attributes could also be the basis for a fairness criterium. E.g. one could demand that algorithms may not use input features of which the correlation with any relevant protected attribute exceeds a certain threshold value. However, the meaning of such a criterium is not evident, since many input features which individually only have a slight correlation to the protected attribute, might be still very indicative for protected group membership in combination. Furthermore, a high correlation between an input feature and a protected attribute cannot show discrimination by proxy.

Alternatively, a qualitative approach could be taken, where for each input feature an argument is made why this feature is or is not likely to cause an unacceptable extent[17] of indirect discrimination. Only if no input features would give rise to such a risk the algorithm would pass this qualitative input feature analysis test. Notice, that this approach is highly subjective and for many features compelling arguments could be made for both sides. Therefore, quantitative feature analyses are poorly suited as audit criteria.

Equality in favourable outcomes

The proof of the pudding is often in the eating. If discrimination is about mistreating people of certain protected groups, we might be less interested in *how* this mistreatment could potentially arise and more interested in *whether* this mistreatment takes place. Hence, there is a clear appeal to algorithmic fairness metrics, which directly consider the impact of an algorithm (or the sociotechnical system it is embedded in) in the form of decision outcomes and show whether the distribution of these outcomes across demographic groups provides evidence of discrimination. However, assessing group fairness from an output perspective is not as straightforward as it might sound. In doing so, again a lot of (implicit) normative assessment choices are faced. Furthermore, any fairness metric implicitly assumes a certain conception of algorithmic fairness. And not all these conceptions might align well with a legal conception of non-discrimination. Hence, when giving an overview of different outcome-oriented fairness metrics, it is important to uncover the conception of fairness that is implicit in each metric.

Before we, continue, we should introduce some key terminology for outcome-oriented fairness metrics. Outcome-oriented (group) fairness metrics are concerned with the distribution of favourable and unfavourable decision outcomes over protected groups. Here, the term *decision outcome* can have two meanings: it can directly refer to the output of a decision-making algorithm (in case of binary classification algorithms), or it can refer to the outcome of the decision-making process an algorithm is involved in. This decision-making process can be considered a sociotechnical system in which humans might make the ultimate decisions, based (at least partly) on algorithmic output. Although both types of decision outcomes could be used in assessing fairness, the important distinction between them, is that when using the former type of decision outcomes, we assess the fairness of (the outcomes of) an

[17] In determining whether a small risk of proxy discrimination can be considered acceptable, quantitative data about how useful/informative the considered feature is in the decision-making process might also be relevant.

algorithm in isolation and when using the latter type, we assess the fairness of (the outcomes of) a sociotechnical decision-making process as a whole. Only when a decision-making process is fully automated (in the sense that an algorithm makes the ultimate decision without human interventions) this distinction disappears.

Next, we should note that to assess group fairness by looking at the distribution of decision outcomes over protected groups, we need a data set that at least contains the decision outcomes and relevant protected attributes of a sample of decision subjects. We will call this data set, the *evaluation (data) set*. For some outcome-oriented fairness metrics, the evaluation set might also need to include the target decision outcomes or input features of all decision subjects it contains. We use the term *subject population* to refer to the population consisting of all individuals who have been subject to a decision-making process within a certain time frame. For now, we assume that an evaluation set is available and that it either contains the whole subject population or a representative sample that is sufficiently large for the purpose of performing reliable outcome-oriented fairness tests. This assumption is implicit in much of the computer science literature on algorithmic fairness but will be challenged later in this chapter. In case we are interested in the decision outcomes of sociotechnical systems instead of algorithms in isolation, we assume that even if the ultimate decision is made by a human, it is still documented and then the outcome-oriented fairness metrics described below could still be used.[18]

Statistical parity

A straightforward way to interpret group fairness is *statistical parity*, also referred to as demographic parity or acceptance rate parity (e.g. Besse et al., 2022; Dwork et al., 2012; Hertweck et al., 2021; Makhlouf et al., 2021; Verma & Rubin, 2018). This fairness measure considers the *acceptance rates* of protected groups. The acceptance rate of a group is the proportion of the members of this group receiving a favourable outcome. According to statistical parity, a decision-making process is fair if the acceptance rate is equal across demographic groups. Hence, if we divide our demographic groups into two protected groups A and B, with acceptance rates R_A and R_B, *statistical parity* demands that $R_A = R_B$. However, exact equality between acceptance rates will never be reached and small differences in acceptance rates can be considered acceptable, so a more realistic demand would be $|R_A - R_B| \leq t$. This demand ensures that the difference between the acceptance rates of both groups, called *statistical parity difference*, does not exceed a certain threshold t. Alternatively, we could demand that $-t_A \leq R_A - R_B \leq t_B$, where t_A is the threshold for how much the acceptance rate for group B is allowed to exceed the acceptance rate for group A and t_B is the threshold for how much the acceptance rate for group A is allowed to exceed the acceptance rate for group B. If group A is privileged and group B is unprivileged, we might care more about preventing negative discrimination of group B than we care about preventing positive discrimination of group B (which would coincide with negative discrimination of group A). This could be achieved by picking a t_B that is lower than t_A.

Another common way of comparing the acceptance rates of the demographic groups is by using the *four-fifths rule*, which attempts to make concept of *disparate impact*, an important concept in USA non-discrimination legislation, explicit (Caton & Haas, 2023; Feldman et al., 2015). This rule states that R_{low}, the acceptance rate of the group with the lowest acceptance rate (often the unprivileged group), may not be lower than four-fifths of R_{high}, the acceptance rate of the group with the highest acceptance rate (often the privileged group). This can be expressed as $R_{low}/R_{high} \geq 0.8$. More generally, for any two protected groups A and B, we could demand that $1/t \geq DIR_{A:B} \geq t$, where $DIR_{A:B} = R_A/R_B$ is

[18] In fact, output-oriented fairness metrics do not even strictly need algorithms to be involved in a decision-making process at all, as long as the required evaluation data set is gather in some way. However, since much of this data is required in training and monitoring ML models anyway, these metrics became popular -and some were even invented- these metrics became popular in context of (ML) algorithms.

the odds ratio of R_A to R_B, which are the acceptance rate of group A and B. We will refer to this odds ratio as the *disparate impact ratio.* Again*,* t is a threshold value, this time denoting the minimal ratio the lower of these acceptance rates should have to the higher. (In case of the four-fifths rule it would be 0.8.) Again, if we want to distinguish between positive and negative discrimination we could also demand that $1/t_B \geq DIR_{A:B} \geq t_A$, where t_A is the minimal ratio from R_A to R_B and t_B is the minimal ratio of R_B to R_A. Hence t_A would be the threshold for negative discrimination against group A and positive discrimination of group B and for t_B this is the other way around. The four-fifths rule has a legal basis in the United States (Greenberg, 1979). However, the translation of disparate impact into the four-fifths rule is disputed (Watkins et al., 2022).Furthermore, within Dutch legal context there are no fixed values for t_A and t_B, as will be shown in chapter 2 of this thesis.

Statistical parity assumes a somewhat radical definition of algorithmic fairness that says that fair treatment means equal outcome across protected groups. Applied to the payroll of a company and the protected attribute sex, this would mean that the average salary of men working at the company should be (roughly[19]) equal to the average salary of women, even if all women working at the company would have fulltime senior and leading positions and all men would be parttime assistants. In other words: statistical parity completely neglects the sound justifications that might explain the difference in outcomes across demographic groups (such as people belonging to different protected groups generally having different functions or working hours).

As discussed above, the goal of promoting group fairness could be either to make sure an algorithm does not treat different people differently or to make society fairer, which might require compensating for existing social inequalities by positive discrimination. Only with this latter aim in mind (or when there are no sound justifications for a difference in outcome across protected groups), statistical parity would be a suitable metric. This is more or less the rationale behind diversity quotas, such as gender quota and racial quota, because of which women or underrepresented races, respectively, are consciously favoured in hiring practices in order increase diversity at the workplace (Shaughnessy et al., 2016).[20]

Conditional statistical parity

Still, there might be many instances in which we do wish to be able to account for sound justifications for differences in treatment across demographic groups. In these cases, it might be preferable to use *conditional statistical parity* instead. This approach allows us to identify a set L of "legitimate" personal attributes, the use of which in the considered decision-making process can be justified convincingly. These attributes should not lead to discrimination, either directly or by proxy. Hence, their predictive value should be based on a causal relationship between the feature and the outcome, without involving protected attributes. Conditional statistical parity is calculated in the same way as regular statistical parity, except that rather than comparing the acceptance rate of two complete demographic groups, we only consider those individuals who have equal values for all features in L (Castelnovo et al., 2020; Corbett-Davies et al., 2017; Verma & Rubin, 2018). This also allows us to calculate conditional statistical parity differences and the conditional disparate impact ratio.

Table 2 shows a hypothetical example that illustrates what the results of a conditional test for statistical parity could look like. The number of combinations for which the test results should be calculated, is

[19] Some difference will be allowed, as long as it is within the thresholds for statistical parity difference or disparate impact ratio.

[20] In fact, assumed that equal numbers of men and women apply for a given job, a gender quota requiring 40 percent of newly hired employees to be woman could be expressed as a minimal threshold of two-thirds on the ratio of the selection rate of women to the selection rate of men.

the product of the numbers of values that each attribute in L could take. The values of attributes that can take a vast number of values (e.g. numerical features) need to be divided into a limited number of categories to ensure that the number of combinations that need to be tested for will be manageable. For example, in tTable 2 the numerical feature gross monthly income is divided into three categories. There is no straightforward and universal method for determining how to categorise attribute values, which is problematic, since the selection of categories can heavily influence the test results.

	Legitimate attributes (L)		**Test results**			
Combination	*Permanent contract*	*Gross monthly income*	*Conditional acceptance rate for group A (R_A)*	*Conditional acceptance rate for group B (R_B)*	*Conditional statistical parity difference ($R_B - R_A$)*	*Conditional disparate impact ratio ($DIR_{A:B}$)*
1	True	< €2,000	0.35	0.41	0.06	0.85
2	True	€2,000 - €4,200	0.72	0.86	0.14	0.84
3	True	> €4,200	0.90	0.98	0.08	0.92
4	False	< €2,000	0.09	0.17	0.08	0.53
5	False	€2,000 - €4,200	0.38	0.47	0.09	0.81
6	False	> €4,200	0.49	0.55	0.06	0.89

Table 2: A hypothetical example of a conditional acceptance rate analysis. *This table serves as an example for what the results of a conditional group fairness test based on acceptance rates could look like for a single division into two protected groups A and B. In this example the set L of legitimate attributes contains the attributes "Permanent contract" and "Gross monthly income". The latter of these features might be a numerical value in the evaluation data, requiring the evaluator to set fixed categories (in this case "< €2,000", "€2,000 - €4,200" and "> €4,200") in order to make a conditional analysis possible.*

Another problem of using conditional measures for equality in outcome is that it is unclear how results such as those presented in Table 2 should be interpreted. Even if we assume that we found reliable ways of choosing between statistical parity and disparate impact and choosing corresponding threshold values, the conditional element of this analysis comes with additional problems. Let us return to Table 2 and assume for argument's sake that the best way to test fairness in this context is to use the four-fifths rule. This rule is satisfied for all combinations in the table except for combination 4. Does this single exception mean that conditional disparate impact is not satisfied? Or should we rather consider the average of all results (0.81) and conclude that conditional disparate impact is satisfied because this number is greater than four-fifths. The problem of interpretation is worsened by the fact that the categorization of features could change both results.

In short, conditional statistical parity (and the derived metrics conditional statistical parity difference and conditional disparate impact ratio) assumes and formalises a conception of fairness that says that treatment is fair if the difference in outcomes between different protected groups can be fully justified by underlying causes that do not relate to the protected attribute that defines these groups. However, this clearly introduces a very impactful normative assessment choice of deciding what input features should count as legitimate reasons for a difference in decision outcomes, independent of the protected attribute. In dividing numerical features into categories and interpreting conditional test results additional normative assessment choices are faced.

Confusion matrix derived parity measures

Equality in chance of getting what one deserves

An alternative conception of group fairness would be that across protected groups, decision subjects should be equally likely to receive the decision outcomes they "deserve". This conception assumes that there exists such a thing like a deserved or a right decision in each decision-making context. This assumption is debatable (and it *will* be debated in the next section of this chapter), but within ML, it is actually quite common to assume not only that such a gold standard for the right output value exists, but also that this standard (or a sufficient approximation of it) is available for training the ML model, in the form of the *target* output value that was introduced in the introduction of this thesis. Assuming, for now, that a right or deserved target decision outcome exists and we have access to a dataset consisting of the input features of many decision subjects, their relevant protected attributes and their target decision outcome as well as the outcome of the decision-making algorithm (or the sociotechnical system it is embedded in) for all of these decision subjects, we can use a set of fairness metrics that (partly) capture this conception of group fairness. To understand these metrics, first we need to introduce the confusion matrix. The confusion matrix, as shown in Table 3, classifies the two types of correct predictions and errors that could be made by an algorithm.

	Actual positive (P)	**Actual negative (N)**
Predicted positive (PP)	True Positive (TP)	False Positive (FP)
Predicted negative (PN)	False Negative (FN)	True Negative (TN)

Table 3: The confusion matrix. *This table shows the names of the two different types of correct predictions (TP and TN) and erroneous predictions (FP and FN) an algorithm could make. The abbreviations P, N, PP, NN, TP, FP, TN and FN are often used to denote the set of all predictions satisfying the corresponding criteria, instead of single instances. This is also how they will be used in this thesis.*

Given our golden standard assumption, false positives and false negatives are cases in which an individual did not get the outcome they deserved. Remember that depending on how we define the classes an algorithm is supposed to classify, a positive classification can either be the favourable or unfavourable outcome. This matters for how we should interpret the different elements of the confusion matrix, as shown in Table 4.

	A positive classification is favourable.	**A positive outcome is unfavourable.**
P	The individual deserves to be advantaged.	The individual deserves to be disadvantaged.
N	The individual does not deserve to be advantaged.	The individual does not deserve to be disadvantaged.
PP	The individual is advantaged.	The individual is disadvantaged.
NN	The individual is denied advantage.	The individual is spared disadvantage.
TP	The individual is deservedly advantaged.	The individual is deservedly disadvantaged.
FP	The individual is undeservedly advantaged.	The individual is undeservedly disadvantaged.
TN	The individual is deservedly denied advantage.	The individual is deservedly spared disadvantage.
FN	The individual is undeservedly denied advantage.	The individual is undeservedly spared disadvantage.

Table 4: The interpretation of confusion matrix elements for two types of decision-making algorithms. *For each confusion matrix class, this table shows the sufficient and necessary condition all individuals in this class satisfy, when a positive classification is favourable and when it is unfavourable. This interpretation assumes that we have access to the decision outcomes each decision subject truly deserves.*

To show how the confusion matrix could be used in assessing fairness, we will consider a hypothetical algorithm used in admitting potential students to a prestigious university programme. In this case, a positive classification (being recommended for the programme) clearly corresponds to the favourable decision outcome. A serious error our algorithm could make, is undeservedly denying students access to the programme, which corresponds to a false negative. It would be clearly desirable to put great effort in minimizing the occurrence of these false negatives over the whole population. However, in view of non-discrimination we are not interested in the quantity of this type of error over the whole population, but rather in the distribution of these errors over different demographic groups. If two out of three members of protected group A (e.g. women) who would deserve admission to the prestigious programme are denied admission, while only one out of nine members of protected group B (e.g. men) who deserve admission is denied admission, this clearly violates the intuitive fairness requirement that protected groups may only be treated differently if this difference in treatment is deserved. What we are comparing for different demographic groups in this example, is the false negative rate (FNR), calculated by dividing the number of false negatives (FN) within a demographic group, by the number of actual positives (P) within this group. Analogous to the case of statistical parity, we could demand *FNR balance* (Chouldechova, 2017; Verma & Rubin, 2018), meaning that $FNR_A = FNR_B$, where FNR_A and FNR_B are the false negative rates of demographic groups A and B defined by a protected attribute, respectively. Again, we must note that absolute equality of these rates will almost never occur. Hence, again we could use the difference or ratio of these rates and decide on thresholds for what difference or ratio we deem acceptable, just like we did with statistical parity.

In the university programme admission example, false positives would be cases in which an individual who would not deserve admission to the university programme is still granted admission. If the probability of being admitted to the prestigious programme while not deserving it, given by the false positive rate (FPR), significantly differs across demographic groups, this also violates the intuition that demographic groups may only be treated differently if this difference in treatment is deserved. Hence, in addition to FNR balance, we could also demand *FPR balance* (Chouldechova, 2017; Verma & Rubin, 2018) or more realistically: we could demand that the difference or ratio of the FPR rates across demographic groups are within certain thresholds. The FPR within a demographic group is calculated by dividing the number of false positives (FP) within this group by the number of actual negatives (N) within this group. In case of algorithms for which a positive classification corresponds to an unfavourable outcome, the meanings of FNR balance and FPR in terms of fairness are swapped. The name *equalized odds* is sometimes used for the requirement that both FNR balance and FPR balance are satisfied (Hardt et al., 2016; Verma & Rubin, 2018).

Equality in chance of deserving what one gets

Comparing FNR and FPR are both ways of testing whether individuals in the evaluation data are equally likely to *get what they deserve*. Alternatively, we could also test whether people across different demographic groups are equally likely to *deserve what they get*. E.g. instead of the chance of someone who deserves admission to a university programme being admitted, we would be interested in the chance of someone who is admitted to this programme, who actually deserves admission. Calculating whether people deserve what they get means calculating both the probability of someone deserving a favourable outcome given they have received this outcome and the probability of someone *not* deserving a favourable outcome given they have *not* received this outcome. The former probability is expressed by the *positive predictive value* (PPV), also called precision. The PPV for a certain demographic group can be calculated by dividing the number of true positives (TP) for this group by the total number of predicted positives (PP) for this group. Likewise, the latter probability is expressed by the *negative predictive value* (NPV), which can be calculated for a demographic group by dividing the number of true negatives (TN) for this group by the total number of predicted negatives (PN) for this group. The term *conditional use accuracy equality* is sometimes to refer to the equality of both the PPV and NPV across demographic groups (Berk et al., 2017; Verma & Rubin, 2018).[21] Again, we could demand that the difference or the ratio of the PPV and/or NPV for two different demographic groups should be within certain thresholds.

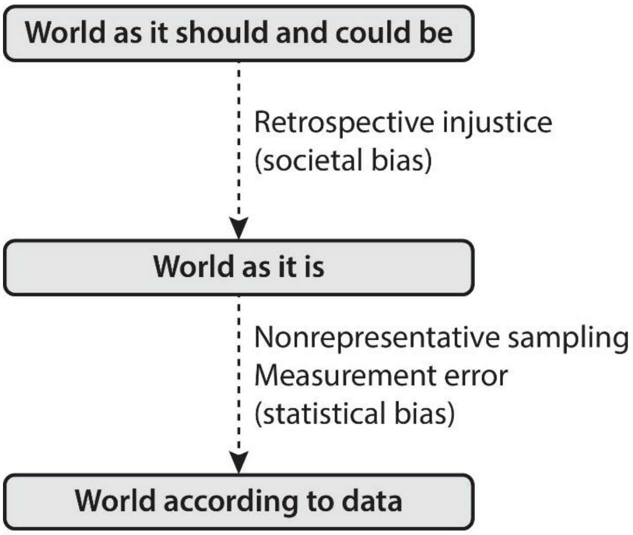
Unfortunately, it has been proven that there is an inherent trade-off in minimizing the difference in PPV between different demographic groups and minimizing the difference in both FNR and FPR between the same groups. Assumed that the share of actual positives is unequal across these groups and our algorithm is imperfect (it makes errors), it will be impossible to simultaneously minimise the differences in PPV, FNR and FPR between these groups (Chouldechova, 2017; Herlitz, 2022). Given the fact that in

[21] Somewhat confusingly, here conditional has a different meaning than in conditional statistical parity.

practice these assumptions will virtually always be met, this trade-off will prevent ML engineers from simply maximizing fairness by minimizing disparity across all possible dimensions and instead they need to choose which conception of fairness to prioritise within the context of their algorithm.

Shortcomings of algorithmic fairness methods

Outcome-oriented fairness metrics do not directly convey information about fairness in the real world but operate solely on a *datafied* version of the world created when collecting the evaluation data. Figure 3 shows how this datafied version of the world differs from the desired, ultimately fair world. Below we will discuss both types of bias included in this figure. Additionally, we will discuss shortcomings in the focus on protected attributes and groups that is inherent in the group fairness approach to algorithmic fairness.



Mitchell S, et al. 2021
Annu. Rev. Stat. Appl. 8:141–63

Figure 3: Two levels of bias in data. *This figure shows two sources of bias in evaluation data ("World according to data"). These biases can invalidate assumptions made by certain fairness metrics. This figure originally appeared in S. Mitchell et al. (2021) and is licensed under CC BY 4.0.*

Statistical bias

In order to train ML models using supervised learning one needs access to a dataset consisting of many datapoints (which in our case represent decision subjects) consisting of all input features used by the algorithm and the target labels. Assuming that an algorithm directly produces decision outcomes,[22] this data set already contains much of the data needed to be included in the evaluation set. This is why the same data gathered when training a model (or more precisely the portion of that data that is used for validating and/or testing), is often also used to validate and/or test its fairness. Of course, this means that ML developers who aim to test for fairness need to ensure that this data contains all relevant protected attributes. An alternative (which is also viable for humanly crafted algorithms and for assessing sociotechnical decision-making processes as a whole) would be to collect evaluation data explicitly for the purpose of evaluating the algorithm (on fairness).

No matter how we get our evaluation data, the fact remains that all fairness tests we perform using it will only tell us whether the algorithm is fair *within the context, or reality, of this data.* However, *statistical bias* will prevent this datafied version of the real world from perfectly representing the real world (S. Mitchell et al., 2021). Statistical bias could be the result of measurement error or nonrepresentative sampling.

In case of a measurement error, the features or target label included in the evaluation data might not represent reality because they are wrongly measured (S. Mitchell et al., 2021). For example, an intelligence score as measured by a certain intelligence test might not truly represent a person's intelligence. Furthermore, this test might be better at detecting intelligence for men as opposed to

[22] These decision outcomes do not need to be the ultimate decision outcomes of this process as a whole for this assumption to hold. All that is needed to assess the fairness of the algorithm itself, is a clear distinction between favourable and unfavourable algorithm outcomes, even if these outcomes might be overruled by a human further down the decision-making process.

women, resulting in unjustified higher intelligence scores for men. If an algorithm using this score satisfied statistical parity, when conditioned on this score, this would seemingly support the conclusion that the algorithm is fair. Apparently, any potential difference in treatment between men and women appears to be explained by a difference in intelligence score, which could be well justified as being a legitimate attribute in many contexts. However, this will not be the case if the measurement of intelligence was not fair in the first place. Alternatively, measurement error affects the target labels, confusion matrix-based fairness metrics will lose their reliability. (More on this below.) Hence, measurement errors are primarily problematic for conditional statistical parity and confusion matrix-based fairness metric.

Non-representative sampling on the other hand, occurs when the evaluation data set is not a representative sample of the total subject population (S. Mitchell et al., 2021). It affects all fairness metrics that rely on evaluation data. We will refer to processes that cause non-representative sampling as filters. For example, it could be the case that to be included in the evaluation data, Dutch citizens must visit government website to give explicit consent, which is explained to them in a Dutch letter. Here the requirement of providing consent is a filter that excludes people who are low-literate (in Dutch) or mistrust the government, many of which might have a migration background. As a result, the group of persons with a migration background in the evaluation data might be small, literate and trusting of the government. Hence, they are a non-representative sample of their group in the actual subject population, meaning that many forms of (indirect) discrimination against people with a migration background that are present in the subject population will not be detected in the evaluation data. It could also be the case that an organisation chooses only to store data about decision subjects in the positive class (those who get selected by the algorithm).[23] Here, the algorithm itself functions as a filter that filters the evaluation data so that it exists solely of positively labelled decision subjects. This also results in the evaluation data being a non-representative sample of the subject population. If the algorithm will be (re-)evaluated on data that it filtered itself, it might enlarge its own initial biases (Kearns & Roth, 2019, Chapter 2).

A straightforward way to solve the problem of unrepresentativeness of evaluation data would be to include the complete subject population (of a certain period) in the evaluation data. However, this would require collecting and storing personal data -including all relevant protected attributes- for all decision subjects, which might be an administrative burden and/or rise privacy concerns.[24] If samples of the actual subject population are used instead, representativeness of the evaluation data should be ensured in another way. If it can be shown that the sampling process is random and unfiltered, this suffices.

There is also a time-related problem concerning the distinction between the population in the evaluation data and the actual subject population. If the evaluation data is representative for the subject population at the start of the lifetime of an algorithm or when an evaluation takes place, this does not necessarily mean it will continue to be representative throughout the further lifetime of this algorithm, creating a new conflict between data and reality. The most rigorous solution to this problem would be to demand the collection of new evaluation data every x years (or months, weeks or days) that is representative for the population subjected to the algorithm during this period. Less rigorous

[23] The EU *GDPR* obligates parties who process data to minimise the amount of personal data they store and process, limiting it to data strictly needed for and relevant to specified purposes. Based on this principle of data minimisation, organisations might decide not to store data about decision subjects in the negative class.

[24] More information about these privacy concerns can be found in the Data and re-evaluation section of next chapter.

(and labour-intensive) alternatives are also imaginable. In work yet to be published[25], Straatman et al. (2023) observe that most of the ethics-based algorithm auditing frameworks do not (explicitly) acknowledge the need to re-evaluate algorithms periodically. In response, the authors proposed a metaphoric periodic appraisal interview for algorithms. Instead of requiring organisations using algorithms to periodically execute a full audit, the proposed framework aims to provide an efficient way of identifying whether an algorithm still functions as intended. This includes assessing whether there was a change in use or context of the algorithm. In a similar vein such an appraisal interview could include questions about whether the data the algorithm's fairness was evaluated on is still representative of the current population of decision subjects. Again, the question whether to opt for a rigorous or more soft way of reassessing fairness, what requirements should hold for showing representativeness of the evaluation data and in this case also the frequency at which reassessments should take place are all matters open for debate.

Societal bias

As stated above, confusion matrix-based balance measures introduce some additional complexity because of their reliance on the target label, which is the "right" or "deserved" outcome of the value the algorithm predicts. In practice, a "right" or "deserved" outcome value often does not exist or cannot easily be found. (E.g. what is the right option when deciding whether to invite an applicant for a job interview based on their CV?). In these cases, there are two ways of "measuring" target outcome values anyway. The first way is by having human domain experts manually label all instances in the evaluation data set based on the available input features. (E.g. all relevant information extracted from a person's CV or simply all text in a CV if we use language comprehension AI models.) The second way is by finding a suitable, objectively measurable indicator for what decision is right. (E.g. we could say that if an applicant gets hired after a job interview, the right decision is indeed to invite them for this interview and if they do not get selected the right decision is not to invite them.) However, since both could be said to be ways of measuring the right decision outcome labels, they could lead to measurement errors, which could in turn lead to discrimination if errors are more often made for people in certain protected groups and to their disadvantage. In case of using indicators specifically, it can be the case that these indicators themselves are not neutral and could thereby be a source of discrimination (Barocas et al., 2023, pp. 34–35). This is illustrated by an investigation by Obermeyer et al. (2019) who showed that using reduction of health costs as an indicator for improvement of health in the USA could lead to racial discrimination, because the amount of money spent in healthcare per black USA citizen is lower than the amount spent per (medically comparable) white citizen. This makes illness of black Americans relatively "cheap" which can cause algorithms trained to minimise health cost to prioritise preventing expensive illness of white Americans, since that is the most effective way of reducing health costs.

Additionally, in many cases in which (ML) algorithms are used in decision-making, the evaluation data consists of historical data of the decision-making process before the algorithm was used. However, before the algorithm was used, the decision-making process was probably not perfectly fair as well, since it was often executed by humans who are prone to have biases as well. Hence, rather than conveying information about how decisions *should have been* made, the evaluation data conveys information about how these decisions *were, in fact, made*. The societal and institutional injustice that causes the difference between the decision outcomes for the whole subject population in an ideal,

[25] I was given insight into this project before publication for the purposes of this thesis. While awaiting official publication, more information about this project can be found on the website of Utrecht University (2024).

optimally fair world and these outcomes in the actual world is referred to as *societal bias* by S. Mitchell et al. (2021).

In effect a combination of societal and statistical bias can often mean that confusion matrix-based fairness metrics cannot reliably be used. Imagine that a decision-making process historically consistently produced less favourable outcomes for women. If an algorithm discriminates women, both the algorithmic output and target labels will show many cases in which women unjustifiably get undesirable decision outcomes (or men unjustifiably get desirable decision outputs), but since we only speak of errors when there is a mismatch between the algorithmic output and target labels these cases will not be counted as errors. Hence it might well be that in the datafied reality of the biased evaluation set the algorithm is fair in terms of any given error rate parity measure while in reality it is much more likely to wrongly provide women with undesirable decision outcomes as opposed to men. In fact, this is closely related to the reason why Amazon's retracted hiring algorithm discriminated against women (Dastin, 2018).

Another problem arises when indicators can only be found for decision subjects who get assigned to the positive class (Lakkaraju et al., 2017). This is the case in the aforementioned example in which being hired after a job interview was seen as an indicator of whether an applicant should be invited to this job interview based on their CV. After all, for those persons who were never invited for a job interview in the first place we can never know whether they would have been hired if they would have been invited. In this case we could compare false positives (people who were invited for a job interview but should not have been invited because they were not hired) across protected groups but comparing false negatives (people who were not invited for a job interview but should have been invited because they would have been hired) is impossible. Hence, if there were many highly competent protected group members who never got invited for a job interview this would remain undetected.

A way to mitigate this problem, is by complementing the selection of the decision-making process with randomly selected persons (Kearns & Roth, 2019, Chapter 2; Wachter et al., 2021). However, in cases where being selected by the decision-making process itself is highly undesirable (e.g. an algorithm that selects people for a thorough and impactful fraud investigation), this might be considered unfair in its own regard (not in terms of non-discrimination, but rather in terms of the right on a consistent decision-making process).

It is important to note that all problems arising from a potential societal bias in target labels (either because they were generated by biased labellers or the way in which they indicate rightness of a decision is biased) only influence confusion matrix-based metrics, which are examples of what Wachter et al. (2021) call *bias preserving* fairness metrics, since they preserve the bias that is present in the evaluation data. Rather than relying on a normative notion of fairness, these metrics take a more descriptive approach, considering an algorithm to be fair if it accurately captures the statistical relations in the (possibly biased) evaluation data. Fairness measures that rely on a definition of fairness in terms of equality in desirable outcomes, on the other hand, are called *bias transforming*. They do not rely on the (potentially) biased target values, but instead offer a new normative rule, such as the rule implied by conditional statistical parity, stating that differences in favourable outcomes across protected groups should only be allowed if they can be explained by different distributions of legitimate attributes across these groups. Since the rightness of a decision can arguably never be translated into data in a truly unbiased manner, the use of confusion matrix-based fairness metrics, should always proceed very carefully.

Problems of protected attributes

In addition to the problems of statistical and societal bias, there are also problems that result from the focus on protected attributes and groups in the group fairness approach to algorithmic fairness. Since discrimination is about a difference in treatment based on a protected attribute, a group fairness analysis should always start with identifying the protected attributes that are relevant potential causes of discrimination in context of the algorithm. Even when a protected attribute is chosen to be included in a fairness analysis there are still decisions that need to be made (S. Mitchell et al., 2021, Chapter 2.2.3). Firstly, we need to decide how we divide our population in protected groups based on a given protected attribute. The attribute sex, for example, is usually used to divide a population into males and females. However, a division into male, female and intersex people might do more justice to the actual biology of all humans. However, since the number of intersex people in any dataset might be quite low, algorithmic fairness test using them as a protected group might not be statistically meaningful. Here the ideal scenario (being able to detect discrimination against as many protected groups as possible) might conflict with the technical reality (the need to have access to sufficient data about all included protected groups).

Furthermore, there is the issue of whether to include intersectional protected groups, which are defined by the possession of several protected attributes. If our protected attributes of interest would be sex (limited to male or female) and sexual orientation (limited to straight or non-straight) the intersectional protected groups would be, straight males, non-straight males, straight females and non-straight females. Intersectionality is important to fairness analyses, since the extent of (algorithmic) discrimination of demographic groups that are intersections of several unprotected groups (e.g. non-straight females) is often larger or at least different than would be expected based on the extent of (algorithmic) discrimination of the unprotected groups contributing to this intersection in isolation (e.g. all females or all non-straight people) (Buolamwini & Gebru, 2018; Cabrera et al., 2019; Escalante et al., 2022; Kim et al., 2020). However, a problem with including intersectional protected groups is that as more protected attributes are stacked in defining them, the total number of groups that should be considered increases rapidly while the number of members of each group decreases just as rapidly. (E.g. There are probably very few to no bisexual, dark-skinned, disabled, conservative, Jewish, Canadian women in most data sets). Hence, rather than including all intersectional groups that can possibly be formed by combining the protected attributes considered in a fairness analysis, it seems more fruitful to select a limited number of intersectional groups to include, based on factors such as relevance to the algorithm's context and size of these groups within the data set. Wang et al. (2022) provide a summary of the problems encountered in intersectional fairness analyses and provide some advice for dealing with them.

Key take-aways for the auditability of algorithmic fairness

Different tests could be executed that might signal discrimination or the absence thereof. However, all measures have their own drawbacks. Table 5 provides a summary of all tests, their uses, shortcomings and the human judgement-based choices involved in performing this test. These tests might be used in establishing audit criteria by requiring a certain outcome of the test (e.g. using a threshold) or by requiring that a test is executed in the first place and its result is documented.

In summary, the most important normative assessment choices faced when assessing algorithmic fairness can be summarised by the following key questions:

Key question I. How should discrimination be defined? Should this definition include or exclude direct discrimination and indirect discrimination, either defined narrowly as strict discrimination by proxy or more broadly?

Key question II. What protected attributes should be used in the assessment? If our definition of discrimination includes indirect discrimination: How should these attributes be used to divide the evaluation data into (potentially intersectional) protected groups?

Key question III. What technical way(s) to test algorithmic fairness should be used? How should the choices specific to the chosen ways to test fairness, be made?

Key question IV. How should the evaluation dataset be obtained? When should de algorithm be re-evaluated on a novel dataset?

In the following chapters we will find guidance in making these decisions in Dutch law and in conversations with practitioners in assessing algorithmic fairness.

	What can it show?	Shortcomings	Choices involved in test
Fairness by unawareness	Whether an algorithm discriminates directly.	Does not show indirect discrimination.	What protected attributes to consider.
Quantitative input feature analysis	Whether an algorithm uses input features that corelate with protected attributes.	Requires access to all protected attributes considered. Correlation between input and sensitive attributes does not necessarily mean proxy discrimination. Indirect discrimination might be caused by combinations of input features. Dependent on (potentially biased) selection of evaluation data.	What protected groups to consider. How to interpret a correlation (by setting a maximum threshold for an acceptable correlation for example).
Qualitative input feature analysis	Whether the person or team executing the analysis deems it likely the input features contain proxies that will lead to indirect discrimination and the argumentation for this.	Highly subjective, therefore unsuited as audit criterium. Proxies might consist of combinations of input features.	What protected groups to consider. How to judge whether an input feature might lead to unacceptable discrimination.
Equality in desirable outcomes	Whether beneficial goods and services are divided equally among demographic groups.	Requires access to all protected attributes considered. Neglects the potentially uneven demographic distribution of certain attributes that legitimise different treatment. Dependent on (potentially biased) selection of evaluation data.	What protected groups to consider. How to interpret inequality (by setting maximal thresholds for acceptable ratios or differences for example).
Conditional equality in desirable outcomes	Whether beneficial goods and services are divided equally among members of demographic groups given that they share certain key characteristics that would otherwise legitimise a difference in treatment.	Requires access to all protected attributes considered. The selection of attributes and categorization of their values are highly subjective. Dependent on (potentially biased) selection of evaluation data.	What protected groups to consider. How to interpret inequality (by setting maximal thresholds for acceptable ratios or differences for example). How to judge whether an input feature is legitimate.
FNR and FPR balance	Whether the proportion of people who get the outcome they deserve (according to the target labels) is balanced across protected groups.	Requires access to all protected attributes considered. Incompatible with PPV balance. Dependent on (potentially biased) selection and labelling of evaluation data.	What protected groups to consider. How to interpret imbalance (by setting maximal thresholds for acceptable ratios or differences for example).
PPV and NPV balance (conditional use accuracy equality)	Whether the proportion of people who (according to the target labels) deserve the outcome they get is balanced across protected groups.	Requires access to all protected attributes considered. Incompatible with either FNR or FPR balance. Dependent on (potentially biased) selection and labelling of evaluation data.	What protected groups to consider. How to interpret imbalance (by setting maximal thresholds for acceptable ratios or differences for example).

Table 5: A summary of different ways to test algorithmic fairness.

Chapter 2: Legal background in the Netherlands

To fully capture algorithmic fairness assessments in an audit framework, all normative assessment choices identified above need to be "eliminated", one of the options for these choices need to be selected. Although it would technically be possible to construct an algorithmic fairness audit framework by arbitrarily eliminating all assessment choices,[26] this would be highly undesirable, since in doing so we would promote an arbitrary conception of fairness as well. Hence, there is need for a non-arbitrary, generally accepted conception of fairness that can eliminate these normative assessment choices, or at least guide them. The current chapter investigates whether this this conception can be found in law, by investigating the relevance of law for the four identified key normative assessment choices. To be more precise, we turn to non-discrimination legislation since this is the area of law has primarily inspired the algorithmic fairness literature. For reasons given in the introduction of this thesis we focus on the Netherlands specifically.

A word of caution is in order here: my own academic education is in Artificial Intelligence, meaning that my attempts to connect the practice of assessing algorithmic fairness to the legal context will inevitably be coloured by my own computer science perspective. I do not have the legal expertise to judge when an algorithm does or does not discriminate. However, I believe that a computer science perspective on law has its own unique value as it might lead to a focus that pays more attention to the technical reality of algorithmic unfairness.

Relevant context

The most obvious place to look for a legal conception of discrimination are constitution articles, regulations and statutes concerned with non-discrimination and equal treatment. However, these sources often leave room for interpretation, partly because it is simply impossible to account for every context in which a law might be appealed to. This is why courts have the authority to interpret law in case of ambiguity or inadequacy. To improve consistency in these interpretations, courts are required to take past judgements on comparable cases (called precedents) into account (Bell, 1997). Because of this, the collection of precedents (in our case on discrimination), referred to as case law, forms an important additional source for finding a legal conception of (non-)discrimination. Given the lack of legal expertise by the author of this thesis, secondary literature on discrimination legislation will play a significant role in this chapter as well. The current section describes the different primary sources in Dutch non-discrimination legislation.

Dutch national non-discrimination legislation

Non-discrimination has a key position in Dutch legislation, at least symbolically, with the very first article of the Dutch constitution (*GW*) declaring: "All persons in the Netherlands shall be treated equally in equal circumstances. Discrimination on the grounds of religion, belief, political opinion, race, sex, disability, sexual orientation or on any other grounds whatsoever shall not be permitted."[27] The non-discrimination article of the Dutch constitution and the additional international non-discrimination directives and treaties the Netherlands is bound to are elaborated upon in the following laws:

- The Dutch Equal Treatment Act (*AWGB*). This is the most generally applicable non-discrimination law in the Netherlands. The *AWGB* prohibits many instances of discrimination

[26] An example of such an arbitrary resolution would be to determine that an algorithm should be considered fair, if it has a PPV ratio between 0.5 and 2 when comparing Christians to Buddhists based on a monthly re-evaluation of the decisions on all decision subjects of that month.

[27] In Dutch: "Allen die zich in Nederland bevinden, worden in gelijke gevallen gelijk behandeld. Discriminatie wegens godsdienst, levensovertuiging, politieke gezindheid, ras, geslacht, handicap, seksuele gerichtheid of op welke grond dan ook, is niet toegestaan."

based on religion, belief, political opinion, race, sex, nationality, heterosexual or homosexual orientation and civil status.

- The Equal Treatment (Disabled and Chronically Ill People) Act (*WGBH/CZ*), which applies to discrimination based on disability and chronical illness.
- The Equal Treatment in Employment (Age Discrimination) Act (*WGBLA*), which prohibits discrimination based on age in the domain of employment and the workplace.
- The Equal Treatment (Men and Women) Act (*WGBMV*), which contains additional articles to protect equal treatment of people of the male and female sex.
- The Working Hours Discrimination Act (*WOA*) which is concerned with equal treatment of people who work full time and people who work part time.
- The Definite and Indefinite Duration Discrimination Act (*WOBOT*) is concerned with equal treatment of people with a permanent and people with a temporary employment contract.

I will refer to this list of laws as Dutch non-discrimination legislation. In practice, when legal cases of alleged discrimination are made, they are based on supposed breaches of these specific non-discrimination laws, rather than the more generally framed article in the *GW*. All Dutch equal treatment acts specify under which circumstances it is prohibited to differentiate between the demographic groups they are concerned with, and which exceptions hold.

EU legislation

As a member state of the European Union (EU), the Netherlands is also bound to EU law, in the form of directives by the European Commission (EC).[28] In fact, EU law takes precedence over Dutch national law when in conflict (Avbelj, 2011; Claes, 2015). However, some of the Dutch non-discrimination laws mentioned above are in fact implementations of EU non-discrimination directives or have been adapted to implement these directives. As a result, all EU non-discrimination directives should be covered by Dutch national law as well (Loof, 2020). Furthermore, national courts can -and in some cases must- appeal to the European Court of Justice (ECJ), which is the supreme court of the EU (European Union, n.d.).

To regulate AI specifically, the EU is currently in the final stages of adopting its highly anticipated *AI act*. A provisional agreement between the Council of the EU and the European Parliament has already been reached and published (Council of the EU, 2023). Key to the *AI act* is its risk-based approach dividing AI systems in categories of low or minimal risk, high risk and unacceptable risk. Unacceptable-risk applications will be banned altogether, and low-risk applications will mostly be subject to voluntary control. High-risk applications however are allowed to enter the European market, under the *AI act*, but only if they conform to specific requirements. AI is considered high-risk when the AI system is an essential component of products that are already covered by other EU consumer protection regulation, such as toys or medical devices, but the *AI act* also defines eight new areas of high-risk AI (*AI act*, article 6 & Annex III).[29] The *AI act* stresses that (high-risk) AI should protect fundamental rights, which includes the right of non-discrimination. Compliance of AI with human rights should be assessed before an algorithm enters the EU market and monitored afterwards. This means that the *AI act* could potentially

[28] These are directives against racial discrimination (*Directive 2000/43/EC*), against discrimination on grounds on religion or belief, disability, age or sexual orientation at work (*Directive 2000/78/EC*) as well as directives against discrimination of men and women in matters of employment and occupation (*Directive 2006/54/EC*), when engaged in an activity in a self-employment capacity (*Directive 2010/41/EU*) and in the access to and supply of goods and services (*Directive 2004/113/EC*).

[29] These are (1) biometric identification; (2) safety components of the management and operation of critical infrastructure; (3) education and vocational training; (4) employment, worker management and access to self-employment; (5) access to essential private and public services and benefits; (6) law enforcement; (7) migration, asylum and border control; (8) administration of justice and the democratic process

lay the groundwork for an audit or assessment infrastructure for algorithmic fairness. However, the final proposal for the *AI act* left much room for interpretation as to how and by which institutions these conformity checks should be carried out (Mökander & Floridi, 2021) and the same holds for the version subject to the recent provisional agreement. Furthermore, the *AI act* still "awaits a formal adoption in an upcoming [European] Parliament plenary session and final Council [of the EU] endorsement." (European Parliament, 2024) and after eventual adaptation, the obligations for high-risk AI will take another 36 months to enter into force (*AI act,* article 85). Hence, it is still too early to tell exactly what this regulation will practically mean for the assessment of algorithmic fairness. This is why this chapter will not go into further detail about the *AI act*.

The General Data Protection Regulation (*GDPR*) is a piece of EU regulation aimed at increasing the control EU citizens have over their personal data and the processing thereof. In contrast to the *AI act* the *GDPR* has been applied since 2018. Several scholars have written about the role of the *GDPR* in enforcing algorithmic fairness or non-discrimination (B. W. Goodman, 2016; Hacker, 2018; Hildebrandt, 2020, Chapter 11.3; Xendis & Senden, 2020). The consensus among these authors is that although the *GDPR* offers little concrete, new rules to improve algorithmic fairness,[30] it does reaffirm that the protected groups that are defined in EU non-discrimination legislation should not be discriminated by applications that depend on data (which includes algorithms) as well. A more important contribution of the *GDPR*, however, is that it prescribes more specific instruments to reinforce compliance with the *GDPR* itself, including its articles about non-discrimination. The *GDPR* offers enforcement instruments of both *ex ante* nature (impact assessments) and *ex post* nature (audits, in a non-strict use of the term).

Enforcement of non-discrimination legislation

In Dutch context, The Netherlands Institute for Human Rights (*College voor de Rechten van de Mens*, CRM) plays a key role in enforcing non-discrimination. The CRM is established by the Netherlands Institute for Human Rights Act (*WCRM*) as an independent institute overseeing human rights. One of their core tasks is ensuring compliance with Dutch non-discrimination legislation. Each Dutch citizen who suspects discrimination can contact the CRM to request a judgement on a discrimination complaint. Judgements by the CRM are not legally binding. However, if the CRM judges an organisation has discriminatory policy, a lawsuit can be filed against this organisation. In this lawsuit the judge is obligated to incorporate the judgement of the CRM in their verdict and since both the researchers of the CRM and the judge are legal experts testing the same facts against the law, they are likely to come to a similar verdict.[31] Moreover, in approximately 70% of all discrimination verdicts by the CRM in which there was an opportunity for the accused organisation to execute structural interventions to stop their discriminatory practices, such action was undertaken indeed.[32]

Regardless of whether they are brought before the ECJ, the Dutch national court or the CRM, legal cases of alleged discrimination consist of two stages. During the first stage, the claimant (a decision subject or advocacy organisation) must provide sufficient evidence to support a presumption of discrimination.[33] The evidence brought forward in this stage has to show that an effect experienced by a person could reasonably be the result of discrimination, but it does not need to prove that

[30] An exception article 9, which will be discussed below.

[31] A difference here is that the CRM only verdicts about discrimination, while a judge might judge that even though discrimination took place, the discriminating party cannot be blamed for this because of justifications that go beyond non-discrimination legislation.

[32] This number can be calculated based on the data on structural measures (Dutch: *Structurele maatregelen*) in table 18 of a report by the CRM (College voor de Rechten van de Mens, 2022a).

[33] In case of a request for judgement brought before the CRM, the CRM will take a more active role in researching whether a presumption of discrimination can be made and in gathering the evidence required for this.

discrimination is the only possible explanation of this effect by excluding all other reasonable explanations of the effect. If the judiciary judges the evidence sufficient to presume discrimination, *prima facie* discrimination is established, and the second stage of the non-discrimination case starts. During this phase, the burden of proof shifts to the defendant (in our case the organisation using a potentially discriminatory algorithm in decision-making). This means that the defendant will be charged guilty of discrimination, unless they are able to provide sufficient evidence to prove that they do not, in fact, discriminate. This two stage process, in which the burden of proof can shift from claimant to defendant, is often referred to as shared burden of proof and is enshrined in the EC non-discrimination directives and the Dutch national non-discrimination laws partly derived from them (College voor de Rechten van de Mens, 2022b; Council of Europe et al., 2018, Chapter 6; Wachter et al., 2020).

Non-discrimination legislation can be enforced both at an individual and institutional level. At an individual level, persons who suspect they have been discriminated can start a case against the alleged discriminator (with or without assistance from advocacy organisations). Conceptually, the black-box nature of many algorithms makes it hard to trace the reasoning behind individual algorithmic decisions, which complicates the support of claims of discrimination by individuals. However, case law shows that if a practice (such as an algorithm) can be shown to be discriminatory to people of a certain protected group in general, this is generally sufficient to establish *prima facie* discrimination against a person who belongs to this protected group and who was disadvantaged by the practice.[34] If an individual claimant wins a discrimination case, among other possible repercussions, the discriminator can be obliged to (financially) compensate the claimant and to end the discriminatory practice. At an institutional level, advocacy organisations can also start a case against a suspected discriminator without an identifiable victim. If the advocacy organisation wins the case, among other possible repercussions, the discriminator can again be obligated to end the discriminatory practice and/or to pay compensation to the advocacy organisation (Veldman, 2021, sec. 4.11). The CRM facilitates individual enforcement of non-discrimination legislation by allowing individual persons to request judgements for cases of alleged discrimination and it can also initiate investigation into more institutional forms of discrimination itself. However, as noted before, any judgement by the CRM is non-binding.

Definition of discrimination

As found in the previous chapter, assessing fairness in terms of non-discrimination requires a definition of discrimination. More specifically, we need to know whether (additionally to direct discrimination) this definition includes indirect discrimination and if so: whether indirect discrimination is defined narrowly as proxy discrimination or whether it is defined more broadly and refers to all cases in which there is a significant difference in treatment between protected groups. This paragraph will show how discrimination is defined in Dutch law.

The *AWGB*, *WGBH/CZ*, *WGBLA* and *WGBMV* specify that the meaning of discrimination[35] includes both direct and indirect discrimination. These laws speak of direct discrimination if a person is treated differently than another person is -or would be- treated in a comparable situation based on the

[34] More on this below in this chapter.

[35] The Dutch word used in these laws is *onderscheid* which can be translated as discrimination but is more commonly translated as distinction. This is in contrast with the *GW* that uses the Dutch word “discriminatie”, the most straightforward translation of discrimination. However, since Dutch non-discrimination law partly serves as an implementation of EU directives that do use the term discrimination, I will translate “onderscheid” as discrimination in this context.

protected attributes covered by the specific law. This legal definition of direct discrimination is compatible with the use of direct discrimination in algorithmic fairness theory.

Notably, there exist proxies which are so strongly or inseparably linked to certain protected attributes that the use of them can be legally considered direct discrimination as well. In this vein, the *AWGB* and WGBMV explicitly state that discrimination on the grounds of pregnancy, childbearing and maternity should be considered as direct discrimination on the grounds of sex. This statement echoes a similar statement in EU legislation (*Directive 2006/54/EC*), which was based on case law of the ECJ. It is unclear when exactly the ECJ considers proxies to be inseparably linked to protected attributes, but it has judged the link between country of birth and ethnic origin not to be inseparable (Weerts et al., 2023). Currently, the three proxies of sex that were mentioned before, are the only proxies for which legislation states that the link between the proxy and the protected attribute is so inseparable that discrimination based on these proxies should be considered direct discrimination.

The *AWGB*, *WGBH/CZ*, *WGBLA* and *WGBMV* speak of indirect discrimination if a seemingly neutral provision, criterion or practice affects persons with a protected attribute specified in this law, in particular, in comparison with other persons without this attribute. This definition of indirect discrimination is an echo of the definition of indirect discrimination used in the EC non-discrimination directives, although these require persons with protected attributes to be put at a particular *disadvantage* instead of being particularly affected, which offers more guidance in interpreting the definition.

Discrimination by proxy, as introduced in the chapter 1, is clearly covered by this definition of indirect discrimination. In this case, the algorithmic decision-making process is a seemingly neutral practice, since it does not directly rely on protected attributes. However, the proxy relationship between input features and a protected attribute causes the algorithm to affect people with in particular if they have this attribute. At first sight, the legal definition of indirect discrimination given here even seems to go beyond a strict definition as discrimination by proxy, since it merely requires protected groups to be particularly disadvantaged by an algorithm, seemingly regardless of whether this difference in outcome is caused by a proxy relationship between the algorithm's input features and a protected attribute. Indeed, to establish *prima facie* discrimination, showing that a protected group is particularly disadvantaged by a provision, criterion or practice suffices, even without showing that this particular specific disadvantage is causally linked to the protected attribute (Tobler, 2008, pt. IV; Wachter et al., 2020).

However, under Dutch (and EU) law, one of the accepted ways for defendants suspected of indirect discrimination to refute established *prima facie* discrimination is by proving to the judiciary that no causal link exists between the protected attribute and the difference in treatment between protected groups (Council of Europe et al., 2018, Chapter 6.1; Wachter et al., 2020). If we assume an algorithm does not use protected attributes as input features, all causal links that could exist between a protected attribute and a difference in treatment between protected groups are caused by a proxy relationship between input features (either in isolation or combination). This leads to the conclusion that ruling out discrimination by proxy for all input features (including their countless combinations) would theoretically suffice to show an algorithm does not indirectly discriminate. Hence, whereas a broad definition of indirect discrimination is used when establishing *prima facie* discrimination, defendants can rely on a narrow definition of indirect discrimination as discrimination by proxy in refuting this claim.

For both direct and indirect discrimination, it is important to realise that in contrast to the *GW*, all other non-discrimination laws and directives mentioned are domain specific. They all specify the domains and contexts in which they apply. Table 6 shows in which domains each Dutch non-discrimination law applies. Additionally, the Dutch non-discrimination laws contain specific exceptions for which making direct or indirect distinctions between people of different protected groups is allowed. These exceptions could apply to specific decision makers[36] or to certain goals achieved by making the distinction, sometimes allowing for positive discrimination.[37] For indirect discrimination specifically, the *AWGB*, *WGBH/CZ*, *WGBLA* and *WGBMV*, as well as all EU non-discrimination directives specify that it is allowed if it is objectively justified by a legitimate aim and the means of achieving this aim are appropriate and necessary.

We can conclude that the Dutch legal definition of discrimination includes both direct and indirect discrimination. Although at face value the legal definition of indirect discrimination only requires a protected group to suffer a particular disadvantage, exceptions to this rule provide reasons to assume the strict conceptualization of indirect discrimination as discrimination by proxy provides a better way to capture the legal definition of indirect discrimination. However, in simply conceptualizing indirect discrimination as discrimination by proxy, one should never forget the legal context of indirect discrimination.

Law	*AWGB*	*WGBH/CZ*	*WGBLA*	*WGBMV*	*WOA*	*WOBOT*
Domain						
Goods and services	Applies	Applies	Does not apply	Does not apply	Does not apply	Does not apply
Workplace	Applies	Applies	Applies	Applies	Applies	Applies
Employment	Applies	Applies	Applies	Does not apply	Does not apply	Does not apply
Liberal profession	Applies	Applies	Applies	Does not apply	Does not apply	Does not apply
membership of labour union	Applies	Applies	Applies	Does not apply	Does not apply	Does not apply
Professional education	Applies	Applies	Applies	Does not apply	Does not apply	Does not apply
Primary and secondary education	Applies	Applies	Does not apply	Does not apply	Does not apply	Does not apply
Social protection	Only based on race	Does not apply	Does not apply	Does not apply	Does not apply	Does not apply
Housing	Does not apply	Applies	Does not apply	Does not apply	Does not apply	Does not apply
Public transport	Does not apply	Applies	Does not apply	Does not apply	Does not apply	Does not apply

Table 6: Application domains of Dutch non-discrimination legislation. *This table contains a summary of all domains on which the elaborated Dutch non-discrimination legislation applies. For each application domain, it shows which Dutch laws fully apply to this domain (the blue cells), which laws do not apply this domain in any way (the orange cells) and which laws do apply to the domain, but only partially (the purple cell). The Dutch constitution (GW) is left out of this table because it describes non-discrimination as a fundamental right that always holds and is not bound to any specific domain. The laws contained in this table partially serve to make this general right more explicit and therefore offer more practical guidance for organisations aiming to prevent illegal discrimination.*

[36] E.g. the *AWGB* contains a clause that renders the law invalid for religious organisations, meaning they are allowed to differentiate between people in ways others are not.

[37] E.g. the *AWGB* contains a clause that allows both direct and indirect distinctions between men and women if the distinction is concerned with protecting women.

Law	*GW*	*AWGB*	*WGBH/CZ*	*WGBLA*	*WGBMV*	*WOA*	*WOBOT*
Attribute							
Sex (and pregnancy, childbearing and maternity)	Applies directly	Applies directly	Does not apply	Does not apply	Applies only for males and females	Does not apply	Does not apply
Gender	Implicit in "any other grounds"	Clarified as form of sexual distinction	Does not apply	Does not apply	Does not apply	Does not apply	Does not apply
Sexual orientation	Applies directly	Applies only for homo- and hetero-sexuals	Does not apply	Does not apply	Does not apply	Does not apply	Does not apply
Civil status	Implicit in "any other grounds"	Applies directly	Does not apply	Does not apply	Does not apply	Does not apply	Does not apply
Race	Applies directly	Applies directly	Does not apply	Does not apply	Does not apply	Does not apply	Does not apply
Nationality	Implicit in "any other grounds"	Applies directly	Does not apply	Does not apply	Does not apply	Does not apply	Does not apply
Age	Implicit in "any other grounds"	Does not apply	Does not apply	Applies directly	Does not apply	Does not apply	Does not apply
Belief	Applies directly	Applies directly	Does not apply	Does not apply	Does not apply	Does not apply	Does not apply
Religion	Applies directly	Applies directly	Does not apply	Does not apply	Does not apply	Does not apply	Does not apply
Political opinion	Applies directly	Applies directly	Does not apply	Does not apply	Does not apply	Does not apply	Does not apply
Disability	Applies directly	Does not apply	Applies directly	Does not apply	Does not apply	Does not apply	Does not apply
Chronical illness	Implicit in "any other grounds"	Does not apply	Applies directly	Does not apply	Does not apply	Does not apply	Does not apply
Working hours	Implicit in "any other grounds"	Does not apply	Does not apply	Does not apply	Does not apply	Applies directly	Does not apply
Permanent/ Temporary contract	Implicit in "any other grounds"	Does not apply	Does not apply	Does not apply	Does not apply	Does not apply	Applies directly
Any ground whatsoever	Applies directly	Does not apply	Does not apply	Does not apply	Does not apply	Does not apply	Does not apply

Table 7: Scope (in terms of protected attributes) of Dutch non-discrimination legislation. *This table contains a summary of all protected attributes included in Dutch legislation. For each protected attribute, it shows which Dutch laws fully apply to this attribute (the blue cells), which laws do not apply this attribute in any way (the orange cells) and which laws do apply to the attribute but either not explicitly or without covering the full diversity of values the protected attribute could take (the purple cells).*

Selection of protected groups

After finding a suitable definition of discrimination, any fairness assessment needs to identify the protected groups that will be included in the assessment. Fortunately, all Dutch non-discrimination laws specify the protected attributes they cover, as summarised in Table 7. For a specific algorithmic application in a given domain, Table 6 can be used to indicate which non-discrimination laws apply to this application, after which Table 7 can be used to find the attributes protected by these laws. Hence, together, these tables can offer guidance for selecting the protected attributes that should be considered in the fairness analysis of a particular algorithm. Of course, these two tables alone cannot convey the true complexity of Dutch non-discrimination legislation. The specific cases in which discrimination is prohibited, the exceptions for which it is allowed and the case law that takes away some of the initial ambiguity in the legislation are all left out of these tables. Furthermore, legislation might evolve over time. For these reasons legal experts should always be involved in identifying the legal boundaries that apply to the use of an algorithm in a specific context. Still, the tables offer an impression of the protected attributes non-discrimination legislation considers.

Once the attributes that are legally protected in the use context of an algorithm have been identified, these attributes should be used to divide the population in the evaluation dataset into different protected groups. In some cases, this is straightforward. When it comes to sexual orientation, the *AWGB* for example, only forbids discrimination based on hetero- or homosexual orientation, meaning that an algorithm subject to the *AWGB* should use the protected attribute sexual orientation to divide the evaluation data into a protected group of heterosexuals and a protected group of homosexuals. (Of course, there will be a considerable remaining group of people who identify as hetero-, nor homosexual, but this group is not protected by the *AWGB* and can therefore be ignored in any fairness assessment that serves only to ensure that an algorithm complies with the *AWGB*.[38])

However, for most protected attributes, it is more ambiguous how they should be used to compose protected groups. This can be because the protected attribute does not lend itself well to be used to divide a population into clear and distinct categories. For example, the protected attribute race translates poorly into a strict division between different protected groups as both the descendance of people and their physical features often associated with a certain race can be too complex to be perfectly captured into a strict division between races. This might require creative solutions, subjective, context dependent classifications or render certain fairness tests unusable.[39]

Even if a protected attribute does, in principle, lend itself well to divide a population into clear and distinct categories, it might be desirable not to use these categories directly as protected groups. Take the protected attribute nationality, for example. Since an algorithm should not discriminate people *of any* nationality, people with the nationality of any nation (officially recognised by The Netherlands) should be their own protected group. In practice, however, it could very well be that there is no good reason to expect an algorithm would discriminate against people from Japan specifically, while not discriminating against people from South Korea (to take an arbitrary example). Therefore, it might make sense to only consider nationalities we suspect might be discriminated against or group nationalities together (in Dutch context this might be Dutch people, EU immigrants and non-EU immigrants). However, (case) law offers no clear guidance for dividing a population into protected groups.

Current Dutch non-discrimination legislation does not explicitly offer any special protection to intersectionally defined protected groups. The *AWGB*, for example, applies to sex and sexual

[38] Any organisation that believes that the exclusion of people who are neither hetero- or homosexual is wrong (as I do) and wishes to make sure these people are not discriminated as well is free to include additional protected groups based on sexual orientation as well.

[39] Still, it is possible to find a workable way to divide a population into protected groups based on race or skin colour as several fairness assessments show (e.g. Buolamwini & Gebru, 2018; Larson et al., 2016). However, fairness assessors should keep in mind that such a categorization will either exclude people of complex, mixed ethnicities altogether or classify them somewhat arbitrarily.

orientation (among other attributes), meaning that men, women, (potentially intersex people), heterosexuals and homosexuals are protected groups within the *AWGB*. However, given the lack of special protection for intersectional groups, this does not make homosexual women, a group defined by combining the protected attributes sex and sexual orientation, a protected group of its own. This means that if an algorithm (or any provision, criterion or practice) subject to the *AWGB* does not particular affect women in general or homosexual people in general, but is found to particularly affect homosexual women, it might not be considered discriminatory under the *AWGB*, since being both homosexual and female is not a separate protected attribute. Within the broader EU legal context intersectionality is not specifically accounted for either (Bullock & Masselot, 2012; Council of Europe et al., 2018, Chapter 2.3; Schiek & Lawson, 2016; Wachter et al., 2020; Xenidis, 2021). The EU non-discrimination case *Parris* even provides juridical precedence in which a claim of alleged intersectional discrimination was considered invalid since there was no discrimination based on any protected attribute in isolation.

In conclusion, with the right legal expertise Dutch non-discrimination legislation can be used to obtain a list of protected attributes relevant to the use context of an algorithm. However, how these protected attributes should be used to divide a population into protected groups is often troublesome and requires making far from straightforward decisions based on somewhat subjective judgements. Currently, Dutch legislation does not clearly forbid discrimination that only affects intersectional protected groups.

How to test for discrimination

When decisions about a suitable definition of discrimination and the protected groups that should be considered have been reached, it is time to decide how to test whether a decision-making algorithm discriminates. Since the legal definition of discrimination includes direct discrimination, performing a *fairness through unawareness* test to rule out the possibility of direct discrimination is sensible. The need of algorithms to satisfy the fairness through unawareness requirement is also affirmed by article 9 of the *GDPR*, which explicitly prohibits "processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, (…) genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation." A prohibition on the processing of these types of data means that they cannot be used as input features for an algorithm as well.[40]

Using statistical evidence to establish *prima facie* indirect discrimination

The question of how to test for indirect discrimination is less straightforward, however. It is generally agreed upon that convincing statistical evidence showing that a rule particularly disadvantages members of a certain protected group is sufficient (but not always necessary[41]) to establish *prima facie* discrimination (Council of Europe et al., 2018, Chapter 6.3; Tobler, 2008, pt. IV; Wachter et al., 2020). However, this information might often not be available to decision subjects who suspect being victim of discrimination. Still, (in the Netherlands) these persons can request a judgement by the CRM, which has a mandate to demand all information and documents reasonably necessary for the fulfilment of its task, which includes judging over discrimination (*WCRM*, article 6). However, to what extend this

[40] B. W. Goodman (2016) considers a minimal and maximal requirement interpretation of this article. The interpretation given here is the minimal requirement. The maximal requirement interpretation is that this article does not only prohibit processing of the protected attributes mentioned in the article but prohibits the processing of proxies of these attributes as well. However, the author admits that satisfying this maximal requirement is "likely infeasible" (B. W. Goodman, 2016, p. 3).

[41] Other indications of discrimination such as common sense arguments that a rule will lead to indirect discrimination can also suffice to establish *prima facie* discrimination.

includes (the data needed to obtain) statistical evidence remains open to debate. (More on this below). Legal consensus on what tests from algorithmic fairness theory are most suitable as statistical evidence would make it easier for organisations to be able to always provide the relevant information and furthermore it will give them a better chance of preventing discrimination before it takes effect. This section aims to find this consensus.

Before investigating the question of which statistical tests from algorithmic fairness theory are best suited to act as statistical evidence, some terminology needs to be introduced. Remember that the legal definition of indirect discrimination is inherently comparative, since it requires a protected group to be particularly disadvantaged *in comparison to* other people. In the following paragraphs we will use the term protected group to denote the specific protected group that is discriminated against according to the claimant. The "other persons" this protected group will be compared to (possibly consisting of all people not in the protected group or all people in the majority or privileged protected group), will be called the *comparator group*. As before, we use the terms favourable and unfavourable group to denote the people who are favoured or disadvantaged by algorithmic decision-making (a practice).

The question of how to use statistical evidence to show that a protected group is particularly disadvantaged by a provision, criterion or practice has been interpreted differently across legal cases. It could be argued that showing a particular disadvantage for a particular protected group merely requires to show that the unfavourable group contains far more protected group members as opposed to comparator group members[42] or similarly, that a large *proportion* of people in the unfavourable group belongs to the protected group (Tobler, 2008, Chapter 2.2). Indeed, in several cases the ECJ, an organisation with a mandate to administer justice on the basis of EU law, accepted evidence of this kind (e.g. *Gester*; *Rinner-Kühn*; *De Weerd*).

However, only considering the proportion of protected group members in the unfavourable group has a serious shortcoming, since such a consideration neglects the proportion of the protected group *over the whole population subject to the contested rule*, even though this proportion might explain the high proportion of protected group members in the unfavourable group. E.g. if there is a rule in place to select aspiring medical students for admission to a medicine programme and around eight out of ten rejected students (the unfavourable group) are women (the protected group), this does not necessarily mean the rule is discriminatory. After all, if around eight out of ten *accepted* students (the favourable group) are also women (meaning that around eight out of ten of all applying students, the subject population, are women), the rule should not be considered discriminatory, since both the distribution of sex among both the favourable and unfavourable group is a fair representation of this distribution over the whole subject population. In fact, the view that the best approach to using statistical evidence should consider both the favourable and the unfavourable group has been explicitly put forward by the ECJ, initially in *Seymour-Smith* to be reaffirmed in later cases (*Villar Láiz*, para 39; *Voß*, para 41).

More specifically, the ECJ states in *Seymour-Smith* (paragraph 59) that:

> "(…) the best approach to the comparison of statistics is to consider, on the one hand, the respective proportions of men in the workforce able to satisfy the requirement of two years' employment under the disputed rule and of those unable to do so, and, on the other, to compare those proportions as regards women in the workforce. It is not sufficient to consider

[42] The demand that the number of protected group members and comparator group members should be similar only makes sense if these numbers are also similar in the population subject to the rule. In many ECJ cases these conditions are met since in these cases the protected group are women, the comparator group are men and the population subject to the rule (e.g. a whole national population) roughly consists of an equal number of men and women. However, one should be careful that these conditions are not always met.

> the number of persons affected, since that depends on the number of working people in the Member State as a whole as well as the percentages of men and women employed in that State."

In *Seymour-Smith*, the protected group under consideration were women and the comparator group were men. However, it seems reasonable that the same standard proposed here would apply to any other protected group and comparator group, since for them as well the reasoning holds that it is not sufficient to consider the number of people affected, since that depends on the number of people subject to the disputed rule and the percentages of comparator group members and protected group members of this total population subject to the rule. A more general version of the standard introduced in *Seymour-Smith* would hence be: the best approach to the comparison of statistics is to consider, on the one hand, the respective proportions of those comparator group members subject to the disputed rule, who are able to satisfy the disputed rule and of those unable to do so, and, on the other, to compare those proportions as regards the protected group members subject to the rule.

Since, in this case, those who satisfy the contested rule make up the favourable group, the proportion of comparator or protected group members subject to a rule, who are able to satisfy this rule, corresponds exactly to what are called the acceptance rates of these groups in chapter 1, which we will here denote as R_C^+ and R_P^+. We will also introduce the concept of a non-acceptance rate of a demographic group, which is the proportion of this group belonging to the unfavourable group. The non-acceptance rates of the comparator and protected group will be denoted as R_C^- and R_P^-.[43] As shown in Appendix A: mathematical details, if everyone in the protected and comparator group either belongs to the favourable group or to the unfavourable group, we can deduce that $R_C^- = 1 - R_C^+$ and $R_P^- = 1 - R_P^+$. The standard proposed in *Seymour-Smith* seems to suggest that both R_C^+ and R_C^- should be compared to both R_P^+ and R_P^-. However, in this same case the ECJ continues by only comparing the acceptance rates R_C^+ and R_P^+ and states that if it can be shown that R_P^+ is "considerably smaller" than R_C^+, this is sufficient to establish *prima facie* discrimination (*Seymour-Smith*, para 63 & 65). Moreover, in *Voß*, paragraph 59 of *Seymour-Smith* is referenced and paraphrased in such a way that a comparison of only R_C^- and R_P^- is required and it is stated that if it can be shown that R_P^- is "considerably higher" than R_C^-, this is sufficient to establish *prima facie* discrimination (*Voß*, para 41-42). As it seems, the ECJ often does not follow its own standard of comparing both R_C^+ and R_C^- to both R_P^+ and R_P^- in discrimination cases (Tobler, 2008; Wachter et al., 2020).

This inconsistency might appear more dramatic than it arguably is. Remember that the objection of the ECJ to only considering the number of people affected by a rule was that *this depends on the number of people subject to the disputed rule (the target population) as well as the percentages of comparator group members and protected group members in this target population*. However, both acceptance rates and non-acceptance rates are proportions and therefore their interpretation does not depend on the total target population size or the percentage of comparator group members and protected group members in the target population. This makes both an individual comparison of R_P^+ to R_C^+ and an individual comparison of R_P^- to R_C^- sufficient to evade the objection raised in paragraph 59 of *Seymour-Smith*, based on which a comparison of both acceptance and non-acceptance rates was proposed. In fact, if (non-)acceptance rates are to be compared by considering their difference, showing that R_P^+ is considerably smaller than R_C^+ is equivalent to showing that the R_P^- is considerably higher than R_C^-, as

[43] In case satisfying a rule is unfavourable to a person (e.g. if the rule is aimed at selecting people for a fraud investigation), the acceptance rate is the proportion of people who do not satisfy the rule and thereby belong to the favourable group and the non-acceptance rate is the proportion of people who do satisfy the rule and thereby belong to the unfavourable group.

proven in Appendix A: mathematical details. However, there is a difference in interpretation of odds ratios when using acceptance rates or non-acceptance rates.

All of this suggests that statistical parity (the demand that $R_C^+ \approx R_P^+$ and $R_C^- \approx R_P^-$) is the most suitable definition of fairness from a legal perspective. However, EU and Dutch law and do not offer guidance on what metrics should be used to show that R_P^+ is considerably smaller than R_C^+ and/or R_P^- is considerably higher than R_C^-. Let alone that legislation offers thresholds for these metrics. As stressed by Wachter et al. (2020), the absence of such threshold values is not a shortcoming of this legislation, but rather a purposeful and advantageous feature. Since context is of key importance in how much inequality can be considered acceptable, positing a universal (context independent) discrimination threshold would be unwise. Still, consensus on what metric to use in the first place would be desirable.

Dutch case law might offer some guidance in finding these metrics and thresholds, as the CRM has experience with using statistical evidence to establish *prima facie* discrimination. They mostly used the (disparate impact) ratio of the non-selection rate of the protected group to the non-selection rate of the comparator group, which can be denoted as $DIR_{P:C}^- = R_P^- / R_C^-$. In the past the CRM (or its legal predecessor to be precise) has even used a fixed upper-bound threshold value of 1,5 for this ratio in cases of alleged sex discrimination[44] (Case number: 2003-91; Case number: 2003-92). Later they started experimenting with more complicated statistical methods developed at Utrecht University. However, this approach has faced criticism because it would make matters unnecessarily complicated and therefore less transparent (Makkonen, 2007, Chapter 3.3).

More recently, the CRM has used statistical evidence to establish *prima facie* discrimination in a case concerning the process that has led to the Dutch Childcare Benefits Scandal, mentioned in the introduction of this thesis. In its investigation, the CRM compared the non-acceptance rates (the proportions of persons flagged as fraudulent or selected for further investigation) for different sub-processes within the bigger process that ultimately could flag persons as fraudulent. In their preliminary study, which was aimed at investigating whether *prima facie* discrimination could be established, the CRM again used the disparate impact ratio $DIR_{P:C}^-$. But in contrast with their earlier rulings, this time the CRM did not explicitly mention any threshold values for this ratio. Instead, it was used to calculate what the proportion of people of the protected group (persons with a non-Dutch background) in the unfavourable group (people who were selected as potentially fraudulent from both the protected and comparator group) would be in the hypothetical situation that exactly half of all decision subjects would be protected group members and the other half would be comparator group members. This proportion can be calculated as $\frac{DIR_{P:C}^-}{1+DIR_{P:C}^-}$ (College voor de Rechten van de Mens, 2022b). If (un)favourable outcomes were completely equally distributed over the population, this proportion should equal 0.5. The closer this proportion gets to 1, the more it supports the suspicion of discrimination against the protected group. For this proportion too no threshold values were given, but this specific study, this proportion was 0.78, from which the CRM concluded that there was a considerable difference in selection rates of the protective and comparator group. Although the rulings of the CRM certainly give some inspiration for how *prima facie* discrimination could be established using statistical evidence, unfortunately, they do not offer a clear and unambiguous picture of this process.

[44] Assuming that women are the protected group and men the comparator group and that a subject population contains an equal number of men and women, this would mean that for every two unfavoured men, there should not be more than three unfavoured women.

Using algorithmic fairness tests to refute *prima facie* indirect discrimination

Once *prima facie* discrimination has been established, it becomes the responsibility of the defendant (the organisation using the algorithm) to proof their algorithm does not, in fact, discriminated by showing that either:

1. There is no causal link between the protected attribute and the difference in treatment between protected groups.
2. The differentiation of the protected group is objectively justified by a legitimate aim and the means of achieving this aim are appropriate and necessary (College voor de Rechten van de Mens, 2022b, Chapter 6; Council of Europe et al., 2018, Chapter 6.1).

The second way to refute established *prima facie* discrimination is of a legal nature and algorithmic fairness tests will be of little help here. The first way, however, could greatly benefit from algorithmic fairness tests.

Remember that the requirement for legal indirect discrimination of a causal link between the protected attribute and the difference in treatment between protected group is exactly the reason why we ultimately believed that the legal meaning of indirect discrimination is better captured by a strict definition of discrimination by proxy instead of a broader definition. This is why fairness tests that are better suited to distinguish justified disparity in decision outcomes from discrimination by proxy, should be used by defendants that wish to refute the claim of discrimination using this first way of doing so. Using a combination of quantitative and qualitative input feature tests to show that each input feature used by the algorithm is not a proxy for the protected attribute, because it either does not correlate with this attribute (quantitative test) or because there are good objective reasons for assuming no proxy relationship (qualitative test) could contribute to support the claim that an algorithm does not discriminate after all. Since the output of an algorithm can only be influenced by its input, one could argue that if there is no proxy relationship between any of the input features and the protected attribute there can be no causal link between the protected attribute and the (difference in) outcomes either. However, using this method it might be hard or impossible to prove that a specific combination of attributes (which are no proxies in isolation) could function as proxies together. Furthermore, to my knowledge there is no precedence of the use of input feature test as evidence in Dutch or EU discrimination cases.

A more straightforward of refuting the causal link between the protected attribute and the difference in outcomes, is by using *conditional* statistical parity metrics. If a defendant can show that the disparity used as statistical evidence to establish *prima facie* discrimination becomes insignificant when conditioning the statistical parity test on a set of legitimate input features, they have strong support for the claim that the initial disparity was the result of a causal link between these legitimate input features and the outcome. Furthermore, if the judiciary agrees that the input features in the legitimate input set themselves are no proxies that could facilitate a causal link between the protected attribute and the differences in outcome anyway, this result also strongly supports the claim that there is no causal link between the protected attribute and the difference in outcome, meaning that the algorithm does not discriminate. This means that there should be no causal pathway of the type shown in figure 2.A between the protected attribute, the decision outcome and any of the features in the set of legitimate input features (both in isolation and combined). In practice, this absence of causality cannot likely be shown with certainty and whether a set of features can be called legitimate indeed, will depend on interpretation. It might be preferable to keep the size of the set of legitimate features small, potentially even one, since this will make it easier to argue that there is no proxy relationship between the members of this set (in isolation and in combination).

In the *prima facie* discrimination study by the CRM, mentioned above, conditional statistical parity metrics were, in fact, included. Several calculations of the odds ratio of the non-acceptance rate of the protected group to the non-acceptance rate of the comparator group were performed, each calculation conditioned on a single attribute such as a person's sex or the number of children they have.[45] In its study report the CRM admits that these conditional tests are not strictly necessary to establish *prima facie* discrimination. Yet, they state these tests provide a preliminary picture of potential alternative explanations for the differences found and they use the finding that conditioning on these attributes did not significantly alleviate the statistical disparity, as additional evidence that supported their establishment of *prima facie* discrimination (College voor de Rechten van de Mens, 2022b). In effect they made the claim of *prima facie* discrimination more robust for the refutation attempts from the defendant after the burden of proof had shifted to their side.

To the best of my knowledge, the active EU and Dutch non-discrimination (case) law does not suggest in any way that if a defendant in an indirect discrimination case can show that the proportions of persons in the protected and comparator group who either get what they deserve or deserve what they get are balanced, the *prima facie* claim of discrimination expires. Hence, there is little to no reason to assume that the use of confusion matrix derived parity measures will produce valuable statistical evidence. This might change in the future, as the provisional agreement on the upcoming EU *AI act* does mention that the technical documentation that should accompany each high-risk algorithm entering the EU market should include information about "the degrees of accuracy for specific persons or groups of persons on which the system is intended to be used" (*AI act*, Annex IV 3).

Thresholds for fairness metrics

The absence of a clear metric and thresholds for (conditional) statistical parity in non-discrimination legislation leaves deployers of algorithms in a challenging position. Relying on only a handful of non-discrimination cases in which statistical evidence was actually considered by the judiciary and even less (seemingly outdates) cases in which an explicit threshold value was mentioned, they have to find a way to interpret the outcome of the statistical parity tests of their algorithms and decide whether an algorithm can be (continued to be) used in practice.

Wachter et al. (2020) propose that beside the significance of a harm (which can be shown by a well-executed statistical parity test), establishing a particular disadvantage for a protected group also depends on the nature of the harm (what is the harm and who does it affect?) and its severity (for each affected person, how severe or damaging is the harm?). These are reasonable factors for deployers of decision-making algorithms to consider in determining how much disparity in decision outcomes they allow. Furthermore, in paragraph 61 of *Seymour-Smith*, the ECJ states that if "the statistical evidence revealed a lesser but persistent and relatively constant disparity over a long period" this can also suffice to establish *prima facie* discrimination, suggesting that the timespan over which an algorithm will operate should also be considered. Finally, case law suggests that societal context is important in judging when we can speak of a particular disadvantage. This was made this explicit in a ruling in a non-discrimination case about a policy that cut large pensions (*YS v. NK*, 2020). Since on average (at least at the time the affected retirees were professionally active) men had higher salaries than women, this policy had more impact on men than on women. However, this generally more negative impact on men was the result of a socio-economical unbalance in favour of those men and hence it was not considered a form of (indirect) discrimination. This example shows that when the group that is allegedly

[45] Whether these attributes are really legitimate is questionable, since number of children could be a proxy for nationality.

discriminated against is privileged, stricter thresholds might be needed to establish *prima facie* discrimination (Weerts et al., 2023).

In conclusion, any deployer of decision-making algorithms who wants to reduce the risk of being found guilty of discrimination would be wise to test for *fairness by unawareness* to exclude the possibility of direct discrimination. Doing a *statistical parity* test (preferably based on disparate impact ratio) and finding that there is no protected group for which the acceptance rate is considerably lower than the acceptance rate of a suitable comparator group is sufficient to prevent a claim of *prima facie* discrimination. If statistical parity is impossible to satisfy (because there are good objective reasons for a different impact on different protected groups), the organisation would be wise to document a *conditional statistical parity* test showing that the unfavourable outcome for a particular protected group can be explained by objective factors (input features) that do not have a proxy relationship with the protected attribute under consideration. If performing a (conditional) statistical parity test is not possible because data on the protected attribute under consideration is not available, a qualitative input feature analysis could be attempted although this is highly subjective. Practitioners involved in assessing algorithmic fairness cannot simply rely on fixed thresholds for the fairness metrics they use, but as Weerts et al., (2023, p. 814, emphasis in original) put it, their focus should shift to "the more difficult yet crucial question of *why* a particular distribution of burdens and benefits is right in a given context, and ultimately, *who* should bear the costs of inequality."

Data and re-evaluation

As discussed, the CRM has the right to demand any information or documents that is reasonably necessary for its judgement in discrimination cases (just as the national court[46]). In principle this could include statistical evidence or at least the data needed to produce this evidence, such as data on the protected attributes of the people affected by the algorithm. However, in the few cases in which such data was used, this use was *ad hoc*. There are no clear guidelines prescribing under which circumstances organisations should be able to provide data that can be used to produce statistical evidence or, perhaps more importantly, what data would suffice. Indeed, the question what information is reasonably necessary for a judgement is very much open to interpretation.

As discussed in the section shortcomings of algorithmic fairness methods (chapter 1), algorithmic fairness tests (including the tests that might serve as statistical evidence) need to be performed on an evaluation data set and in practice, in case of ML algorithms, the evaluation data is often the same data used to test the performance of the ML model when training it. This is often historical data of the decision-making process before the algorithm was introduced. This means that this data does not show the actual effects of using the algorithm on real persons, but rather the hypothetical effect the use of the algorithm would have had if it would have been used in the cases contained in the evaluation data set. Hence, the use of evaluation data in legal contexts raises many questions. Is it sufficient for a defendant in a discrimination case to provide the historical data an algorithm was initially evaluated on and its outputs for this data? If the fairness tests that can be applied to this data do not show a (hypothetical) particular disadvantage for the protected group under consideration, is this enough to close the case (if the claimant does not provide other sufficient evidence to establish *prima facie* discrimination at least)? Or does the defendant need to provide data about the actual impact the algorithm had after its deployment? And if so: does this data need to be about the specific period in time during which the suspected discrimination took place and how should (the length of) this period be determined?

[46] This is enshrined in article 22 of the Civil procedure code of the Netherlands (*RV*).

In answering each of these questions, non-discrimination (case) law seems to be of little help. One could argue, however, that, data about the actual impact of the algorithm should preferably be used in producing statistical evidence, since the legal definition of indirect discrimination is concerned with the disadvantageous effect of a seemingly neutral rule on a protected group. Furthermore, this would suggest that the data used for this purpose should be as reflective as possible of the context in which the supposed discrimination took place, meaning that data from a period close to the period in which this took place is preferable. This interpretation of the legal definition of indirect discrimination seems to be in favour of frequent (re-)evaluations of an algorithm after it has been deployed.

Furthermore, as mentioned before article 9 of the *GDPR* complicates and potentially completely prevents processing certain forms of personal data. This includes data related to race, belief, religion, political opinion, disability, chronical illness[47] and sexual orientation (van Bekkum & Zuiderveen Borgesius, 2023), all of which are protected attributes in certain domains in Dutch non-discrimination legislation. This does not only prevent these attributes from being used as input features (thereby preventing direct discrimination), but it also prevents these attributes from being used in algorithmic fairness assessments. Future legislation might change this, article 10 (5) of the provisional agreement on the EU *AI act* does (under strict conditions) explicitly allow for the processing of said attributes if "strictly necessary for the purposes of ensuring bias monitoring, detection and correction in relation to the high-risk AI systems." Yet, as long as the *AI act* has not been formally adapted and did not fully enter into force, this data cannot be collected, except when certain exceptions within the *GDPR* itself apply, which often will not be the case (van Bekkum & Zuiderveen Borgesius, 2023). Of course, the judiciary cannot demand a defendant to provide data which contains protected attributes the defendant is not allowed to store in the first place, meaning that fairness tests (in their standard form) might be limited to those protected attributes the *GDPR* does not forbid collecting such as sex and age. In the Chapter 3: Assessing algorithmic fairness in practice chapter ways of adapting fairness tests when protected attributes are not available, will be discussed.

Key take-aways for the auditability of algorithmic fairness

This chapter aimed to answer the question of how an algorithmic fairness audit should be executed given the aim of complying with Dutch non-discrimination legislation. More specifically, to the extent to which this is possible, it aimed to answer the normative assessment choices identified in chapter 1 from a legal perspective. The questions of how discrimination should be defined, and which protected attribute should be considered in the fairness assessment could be answered relatively well from a legal perspective. The question of how to use these attributes to compose a protected group was less easily answered. The legal guidance in finding a specific way to test fairness is somewhat ambiguous as is the guidance into which data to use for evaluation and when to re-evaluate. In its current state, the way non-discrimination legislation should be applied on algorithmic decision-making is highly ambiguous, leading to a risk of algorithmic discrimination being not well addressed by the judiciary and organisations using algorithms in decision-making not knowing how to ensure compliance with non-discrimination legislation. The most important conclusion that can be drawn from this chapter is that although Dutch legislation and regulation does forbid algorithmic decision-making processes (just as any other decision-making processes) to exhibit many forms of direct and indirect discrimination, it does not always provide the details necessary for organisations using algorithms in their decision-making processes to reliably prevent this discrimination, let alone that the legal conception of non-discrimination is clear enough to be translated into objective audit criteria. Partly, the lack of these

[47] Article 9 (1) of the *GDPR* prohibits processing of "data concerning health", which is taken to include data on disability and chronical illness.

details might be deliberate, since leaving room for the contextual interpretation essential to non-discrimination law, requires leaving details open.

Below, the questions raised in the extent to which this is possible, it aimed to answer the non-technical questions raised in chapter 1 and the best answers a legal context can offer are summarised.

Key question I: *How should discrimination be defined? Should this definition include or exclude direct discrimination and indirect discrimination, either defined narrowly as strict discrimination by proxy or more broadly?*

The legal definition of discrimination includes both direct and indirect discrimination. The legal definition of direct discrimination aligns well with the use of the term in algorithmic fairness theory. The legal use of indirect discrimination in establishing *prima facie* discrimination merely requires a protected group to be particularly disadvantaged without demanding this disadvantage to be causally linked to the protected attribute, but since *prima facie* discrimination can be refuted by showing that there is no causal link between the protected attribute and the unfavourable outcome for the protected group, a more strict definition of indirect discrimination as discrimination by proxy seems more suitable. Legal exceptions might also make discrimination by proxy or sometimes even direct discrimination is admissible.

Key question II: *What protected attributes should be used in the assessment? If our definition of discrimination includes indirect discrimination: How should these attributes be used to divide the evaluation data into (potentially intersectional) protected groups?*

The protected attributes that should legally be considered in assessing the fairness of an algorithm can be found by first identifying the non-discrimination laws that apply to the context of the algorithm and then identifying the attributes protected by these laws. Table 6 and Table 7 could guide this process. There is no clear legal obligation to specifically protect intersectional demographic groups. Composing protected groups based on many protected attributes can be challenging and remains dependent on contextual judgement.

Key question III: *What technical way(s) to test algorithmic fairness should be used? How should the choices specific to the chosen ways to test fairness, be made?*

To assure an algorithm does not discriminate directly, a fairness by unawareness test appears sufficient. To prevent the establishment of *prima facie* indirect discrimination against a relevant protected group, an organisation should be able to show that the selection rate of this protected group is not considerably lower than the selection rate of a(ny) suitable comparator group and/or that the non-selection rate of this protected group is not considerably higher than the non-selection rate of this comparator group. This is what was called statistical parity in chapter 1. Although there are few examples of actual calculations used to identify such a considerable difference, a study by the CRM used to establish *prima facie* discrimination suggests that using (a metric derived from) the odds ratios of non-acceptance rates is a suitable method. Although a combination of a quantitative and qualitative feature analysis might theoretically serve as evidence to refute *prima facie* discrimination once established, a more straightforward way of refuting *prima facie* discrimination using tests from algorithmic fairness theory is by showing that the considerable difference used as evidence for *prima facie* discrimination disappears when conditioning on a legitimate attribute or set of legitimate attributes. However, this requires that each of these attributes certainly does not facilitate a causal link between the protected attribute and the difference by serving as proxy (in other words: conditional statistical parity is satisfied). The study by the CRM again suggests using odds ratios of non-acceptance rates is a suitable method. Dutch legislation does not unambiguously offer specific thresholds for "considerable differences" in acceptance rates, as the acceptable difference might be context dependent.

Key question IV: *How should the evaluation dataset be obtained? When should de algorithm be re-evaluated on a novel dataset?*

It could be argued that for any case of suspected algorithmic discrimination, an organisation should be able to provide data that is representative of the period in which the alleged discrimination has occurred, which would suggest that pretesting fairness during algorithm development does not suffice to comply with non-discrimination legislation. This would mean that organisations wanting to make sure their algorithms comply with the law should frequently re-evaluate the fairness of their algorithm on recent evaluation data. However, this interpretation of the law is poorly grounded in (case) law or algorithm specific guidelines. Furthermore, privacy regulation might limit the data that could be collected and therefore could be demanded for legal purposes.

Chapter 3: Assessing algorithmic fairness in practice

In chapter 1, we identified four main normative assessment choices, that should be made to perform an algorithmic fairness assessment. In chapter 2, we found that these choices are not fully eliminated by Dutch non-discrimination legislation. From this, we can already conclude that it is not possible to fully capture algorithmic fairness into audit criteria, such that meeting these criteria guarantees non-discrimination. Given the impossibility of using audits to directly assure non-discrimination, we will explore alternative possibilities for using audits with the aim of preventing non-discrimination in more indirect ways. Instead of (fully) relying on law, these alternative forms of audits could also incorporate best practices that can be identified in the algorithmic fairness assessment practices that are already performed by organisations (without them being mandated or following a strict audit framework). To explore what these best practices could look like and how practitioners performing these assessments deal with the normative assessment choices they face, this chapter connects this thesis to the actual practice of assessing algorithmic fairness within organisations.

Methods to gain insight into the practice of assessing algorithmic fairness

To explore best practices in assessing algorithmic fairness, I needed a way into the practice of assessing fairness. Several methods to gain this access were considered. The initial plan was to join Utrecht University's Data School in their Fundamental Rights and Algorithms Impact Assessment (FRAIA) sessions.[48] The Data School is a research platform which is partly focussed on responsible AI and data in practice. Commissioned by the Dutch Ministry of the Interior and Kingdom they developed FRAIA as an assessment framework, which organisations can use to check whether their (intended) use of algorithms aligns with the fundamental rights contained in Dutch law. This framework has the form of a document that guides organisations through a list of questions on the consideration of elements relevant to human rights for various aspects of algorithms (Gerards et al., 2022). The Data School helps governmental institutions in applying FRAIA by guiding them through the application of it on algorithms that are (intended to be) used by the organisation. In the past, the Data School has used the guidance sessions of their more general data ethics tool DEDA (Franzke et al., 2021) as an opportunity to make fieldnotes during those sessions and engage in participatory observation which they can use to gain unique access to the discourses on data ethics within the (governmental) institutions they guided (Siffels et al. 2022; Schäfer et al., 2023). Having cooperated with the Data School myself, I can confirm that a similar approach was used in the FRAIA sessions during which fieldnotes were made as well.

The plan was that this thesis would use participatory observation facilitated by FRAIA and the guidance sessions of the Data School to investigate the practice of assessing fairness. However, this plan was abandoned, mainly because it did not yield the type of information needed to determine how fairness is assessed in practice. On first sight FRAIA sessions do seem a good place to gain insight in fairness assessments in practice. The fundamental rights, which FRAIA is designed to protect, include equality and non-discrimination rights and FRAIA does contain questions about:

- "measures that can be taken to counteract the risks of reproduction or even amplification of biases" (Gerards et al., 2022, p. 39),
- the "assumptions and biases [that] are embedded in the data" (Gerards et al., 2022, p. 28) and the methods to overcome and mitigate them and
- the risk of "stigmatization, discrimination or otherwise harmful or adverse effects on citizens" (Gerards et al., 2022, p. 53) and the methods to combat or mitigate them.

[48] Dutch: *Impact Assessment Mensenrechten en Algoritmes (IAMA)*

Together, these questions could be taken as an invitation to perform and document a thorough assessment on both the fairness of the outcomes of the algorithm and the biases embedded in the training and evaluation data. However, whether this invitation is accepted depends on the motivation of the organisation performing the assessment to delve into the best ways to use fairness assessments to reduce the risks of algorithmic discrimination. FRAIA itself offers little guidance here, except some references to other sources that might be helpful. Additionally, not all algorithms (in a broader sense of the word) used by government institutions are used in decision-making processes and therefore relevant to this thesis.

An additional problem is that the assessments completed by the organisations the Data School guided, were not shared with them. This means that for the FRAIA sessions that had taken place before the start of the research phase of this thesis, I could only use the data contained in the fieldnotes that were made by the Data School employees for my thesis. However, since these fieldnotes were not written with a focus on algorithmic fairness in mind, they could not be used for our current research aims. This means that I could only get relevant data from FRAIA sessions by attending them myself and making my own fieldnotes with a focus on algorithmic fairness. However, of the FRAIA sessions I would have been able to attend during the research phase of my thesis, there were too few with a strong risk of discrimination to consider participatory observation of these sessions a suitable method for gaining the desired access to the practice of assessing algorithmic fairness.

An alternative plan was developed in the form of a qualitative comparative analysis (QCA) of algorithmic fairness assessments (including audits). A QCA is a valuable method for discovering whether the lessons drawn from algorithmic fairness theory and legal context are in practice incorporated into algorithmic fairness assessments and what the result of fairness assessments might ultimately look like. Unfortunately, however, fairness assessments and audits that have been publicly shared and hence available to me to analyse, are scarce. Furthermore, this type of analysis only provides insight in the final reports that resulted from the fairness tests. Not all deliberations leading to important, normative decisions in approaching algorithmic fairness (e.g. what fairness definition(s) are used) are included in these reports. It is exactly those deliberations that are useful in finding alternative ways of using audits in ensuring algorithmic fairness.

Chosen method: informal conversations with practitioners

Given the inadequacy of the alternatives, it was finally decided to use informal conversations as main method for gaining insight into the practice of assessing algorithmic fairness. These conversations were held with four professionals with different backgrounds involved in assessing algorithmic fairness. Being able to talk with these practitioners themselves, provides invaluable insight into the deliberations behind a fairness assessment well beyond the plain results that might and up in a published report (if a report will be published in the first place). Furthermore, the informal nature of these conversations and the anonymization of the excerpts that would be used in this report, allowed the practitioners to speak more freely about their experiences in assessing algorithmic fairness. The conversations were semi-structured, using a list of questions to fall back on, but allowing the conversations to flow naturally and the practitioners to provide their own input.

In the conversations, I focused on which methods from algorithmic fairness theory the practitioners had used and considered and why they did or did not use certain methods. Like chapter 2, this chapter will be structured around the four main questions on normative assessment choices that were identified in chapter 1, since those are the open questions that best practices could provide an answer to. The sample size of practitioners spoken to, is too low to make generalisations and hence the aim of this chapter is not to identify any definite commonly accepted best practice. Instead, the aim of this chapter is to explore what best practices *could* entail. For this explorative aim, the chosen qualitative approach is particularly suited, since it enables us to go more in depth on the content of potential best practices, enabling us to shape and define them, instead of focussing on whether a predefined potential best practice is followed commonly. Given the relative novelty of algorithmic fairness assessments and the scarceness of publicly shared methods and results, I would argue that we are still in the exploration phase of identifying best practices, which is why a qualitative, explorative research method was chosen. Costanza-Chock et al. (2022) took a similar approach in which they first conducted semi-structured interviews with a small sample (n=10) of leaders in the field to get qualitative insight in emerging best practices in algorithmic auditing, after which they conducted a largescale survey among people involved in algorithmic auditing (n=152). Our research in this chapter adds to theirs by being specifically situated in the Netherlands, thereby being relevant to an EU context (whereas most of their respondents and interviewees were from the USA) and by being specifically focused on algorithmic *fairness* assessments, instead of algorithmic audits (or assessments) in general.[49] Constituting only one of three chapters in an individually executed master thesis, the research in this chapter will be limited to the exploratory, qualitative phase, leaving the quantitative phase (focused on algorithmic fairness audits and Dutch or EU context) for future research. However, this chapter will often refer to the research by Costanza-Chock et al. (2022) to give some idea of whether the findings in my conversations are more widely supported.

The conversations were held in Dutch and were recorded. The excerpts of these recordings, which are relevant to the current analysis were first transcribed in Dutch and then translated to English. The four practitioners included in the research will be identified as person A, person B, person C and person D, to provide anonymity. The excerpts used in this chapter are referred to by the letter of the person who said it, followed by a dot and a number. They can be found in appendix B of this thesis in order of appearance in this chapter. Comments that serve to reduce ambiguity, provide additional explanations were deemed helpful or improve correct use of language were added to the excerpts between square brackets. When excerpts skip part of a recording because this part contains irrelevant information, duplicate information or interruptions by the researcher this is indicated by "(…)". To further provide anonymity of the practitioners spoken to, we will use the gender-neutral pronouns they/them to refer to them. Furthermore, the organisations for which they work will not be named and if they were named in the interview excerpts included in this chapter, those names will be replaced with "[name of organisation]".

For this research, I spoke:

Person A: The lead data scientist at a private "Fortune 500" company involved in recruitment. The conversation was about an internal fairness assessment.

Person B: A data scientist working for a major Dutch municipality. The conversation was about an internal fairness assessment.

Person C: A board member of an organisation that executes external fairness assessments. The conversation was about a third party, external fairness assessment in the public sector.

Person D: A consultant at a large auditing/consultancy firm. The conversation was about their guidance in two internal fairness analyses in the public sector.

The most important criterium for selecting practitioners to be included in the research was the requirement of having been personally involved in at least one fairness assessment, using at least one of the methods described in the Chapter 1: Algorithmic Fairness chapter. Furthermore, in order to ensure diversity in perspectives, the sample of practitioners included in the research had to include

[49] Costanza-Chock et al. (2022) use the term "audit" in a non-strict manner, meaning that the algorithmic fairness assessments covered in this chapter would likely satisfy their definition of audits as well.

practitioners from both the private and public sector with different relations to the organisations deploying or planning to deploy the algorithm under consideration. Additionally, people with senior or leading roles in their organisations or teams were preferred, following the assumption that they have most knowledge about the algorithm and/or its assessments. Table 8 shows how these attributes were distributed over the sample of practitioners included in the research. The private sector is slightly underrepresented in the sample. This might reflect the apparent fact that the interest into algorithmic fairness is greater in the public sector, as the teams of both person C and person D were in principle open to working with both (the cases of) public and private organisations, but private organisations did not reach out to the team of person D and their use cases were not available to the team of person C.

It should be noted that person D has cooperated with person B at the same municipality before person D started at a consultancy firm. Hence their experiences might not be truly independent. In my conversation with person D, I did however focus on their work at the consultancy firm.

Person	Role	Relation to auditee	Sector of auditee(s)
A	Lead data scientist	Internal (first party)	Private sector
B	Senior data scientist	Internal (first party)	Public sector
C	Board member of auditor	External (third party)	Public sector
D	Senior consultant	External (consultancy)	Public sector

Table 8: Summary of the practitioners included in the research. *In this table, the word auditee is used to refer to the organisation that deployed or planned to deploy the algorithm that was subject to the fairness assessment. It is not meant to imply that these assessments were audits in the narrow definition of the word.*

Findings on the definition of discrimination

In the chapter 1, it was found that the concept of discrimination can be subdivided into direct and indirect discrimination, where indirect discrimination can either be defined loosely as a difference in outcome between protected groups or more strictly, requiring the difference in outcome to be the result of the protected attribute that was inferred through the proxy. In chapter 2 it was found that, in principle, non-discrimination legislation applies to both direct and indirect discrimination, where a loose definition of indirect discrimination should suffice to establish *prima facie* indirect discrimination, whereas the defendant can refute this *prima facie* claim by showing that, defined in its stricter sense, no indirect discrimination has occurred.

Unfortunately, during the time these conversations were held, my findings on the different ways in which discrimination could be defined were not complete yet. Hence, it did not occur to me to ask the practitioners about the definition of discrimination they used in their fairness assessments. It can be deduced from my conversations, however, that none of the practitioners spoken with considered the elimination of sources of direct discrimination sufficient to consider an algorithm fair, since all of them included tests that can detect indirect discrimination in their fairness assessments.

It should be noted that it is quite possible that (some of) the practitioners did not explicitly define discrimination before commencing their fairness assessments. This can easily happen if a fairness assessment is primarily approached as a technical exercise instead of a legal instrument. In that case, the question of what improving fairness or reducing discriminatory bias means will only be faced once a specific way to measure (a specific conception of) fairness is chosen (more on that below).

Findings on protected attributes and groups

In the chapter 1, we established that many fairness metrics require access to evaluation data that can be divided in protected groups based on protected attributes. In chapter 2 we saw that non-discrimination law can serve as a source for finding these protected attributes, although using these attributes to divide a population in protected groups is not always straightforward. Hence, it is interesting to see whether the practitioners included in this research based their choice of considered protected attributes on law indeed and how they approached the problem of using these attributes to construct protected groups.

Persons A, B and D all mentioned a conflict between what protected groups should ideally be considered in a fairness assessment and what protected groups could be considered in reality (A.1; B.1; D.1). To determine what protected groups should ideally be considered different methods were used persons A and D mention reliance on the law (A.1; D.1). This confirms the earlier conclusion that the law could be a source of finding protected attributes and that this is used eventually, in practice. Person B mentions consulting business stakeholders and doing a preliminary data analysis (B.1) and person D additionally mentions consulting prominent publications and finding best practices (D.1). All three practitioners admitted that (a lack of) data availability was the reason why, in reality, some protected groups that should ideally be considered could not be considered in practice (A.1; B.1; D.1).

Dealing with unavailable data

Several methods that could potentially help to work around the data availability problem were mentioned in the conversations. Person C, for example, took a different approach to testing algorithmic fairness that did not rely on availability of data on protected attributes. They performed an external fairness assessment of an algorithm that had already been scrutinised by journalists, who had already analysed which (intersectional) groups had been disadvantaged by the algorithm. Therefore, the team of person C approached their assessment by starting from the input features and reasoning how these could potentially lead to discrimination (C.1). Person D also mentions use of this method (D.2) This approach corresponds to the qualitative input feature analysis that was introduced in chapter 1. The advantage of this approach is that it only requires access to the input features that were used and no access to the model outcomes or target labels, thereby avoiding the data availability problem mentioned above.

Another option for bypassing the data availability problem is by identifying proxies of the unavailable protected attributes and basing the fairness test on those proxies. In this context, any input feature causally or statistically linked to a protected attribute is considered a proxy for this attribute.[50] ZIP code, for example, is a well-known proxy for race, so in absence of direct data on race, one could form a protected group of people with ZIP codes most clearly associated with a certain race and test whether these people are particularly disadvantaged compared to people with other ZIP codes. All methods listed in chapter 1, relying on access to a protected attribute, can also be executed using proxies for these attributes. Methods for using proxies in fairness assessment have been proposed by scholars (e.g. Chen et al., 2019; Galhotra et al., 2021; Kallus et al., 2021; Zhu et al., 2023) and used in practice by both X (then known as Twitter) (Belli et al., 2023) and Meta/Facebook (Alao et al., 2021).

Indeed, both person B and person D mentioned the use of proxies in this way (B.2; B.3; D.3). Both excerpt B.3 and excerpt D.3 show that a downside of using single proxies to stand in for unavailable protected attributes, is the interpretation of the results that are gathered this way. Only if a proxy is very strong, it could be said that differentiation based on this proxy means discrimination based on the underlying protected attributes, person B and D think. Additionally, person B is worried that in approximating sensitive attributes about individuals one ignores the fact that this information is protected by privacy regulation for a good reason (B.2). The aforementioned study by Meta/Facebook tries to overcome this privacy concern of approaching protected attributes using proxies by only assessing fairness on an aggregated group level and not predicting the protected attributes (in this case

[50] The need for the predictive value of a proxy stemming from its relationship to a protected attribute is only relevant in context of discrimination by proxy and can be ignored here.

race) on an individual level (Alao et al., 2021) and Zhu et al. (2023) claim that using the algorithm they propose there is no need for a strong proxy relationship in order to assess fairness and furthermore, using weak proxies instead would also serve to protect privacy.

Intersectionality

As mentioned in chapter 1, people with multiple unprivileged protected attributes can be particularly disadvantaged by algorithms, due to a phenomenon known as intersectionality. This can be a reason to include intersectional protected groups (protected groups defined by a specific combination of multiple protected attributes) into a fairness assessment. This is not clearly demanded by (Dutch) non-discrimination legislation, however.

Persons A, B and D held different attitudes towards including intersectional protected groups. Both person A and person D think that intersectionality is something that should be considered when assessing algorithmic fairness (A.2; D.2). Yet, person A did not actually account for intersectionality in practice (A.2) and for person D it is unclear whether they did (D.2). Person B also did not test for intersectional effects and does not think they should have, for two reasons: firstly, the domain or business experts never indicated a risk of intersectional discrimination and secondly, including intersectional protected groups would introduce a new problem of data availability, since the size of protected groups can quickly decrease when more protected attributes are used to define them (B.4). The data availability problem is also acknowledged by person D, whereas the argument of domain experts not indicating the necessity of intersectionality stands in contrast with the experience of person D that domain experts often do point at the risk of disadvantaging intersectional groups (D.2). This could have various explanations such as the domain experts consulted by person B and person D having different backgrounds or the use cases both persons were involved in simply being different. Costanza-Chock et al. (2022) found that 65% of their respondents reported to conduct intersectional fairness assessments. Yet, the authors questioned whether all of them did this in practice given the data availability problem.

In conclusion, this section shows a friction between ideals and practice. Data availability greatly limits what relevant protected attributes found in legislation or elsewhere can be used for fairness tests in practice. To overcome this problem, one could focus on the input instead of the (differences in) output, by doing qualitative input feature assessments to identify (potential) proxies for protected attributes that are not in the data. These proxies can be simply removed from the input features to reduce the risk of indirect discrimination and they can also be used to divide the evaluation data into groups instead of using the unavailable protected attributes themselves. Including intersectional protected groups in fairness assessments is hard, primarily because it creates a new data availability problem since the size of protected groups decreases as the number of protected attributes used to define them increases.

Findings on testing algorithmic fairness

As concluded in chapter 1, even after deciding which protected groups to compare in an algorithmic fairness assessment, there are multiple ways of executing such a test and different fairness metrics could be used. Chapter 2 showed some precedents for using statistical parity-based metrics in judicial context, although much about the use of which metric exactly and the interpretation of its outcome remained unspecified.

The persons in our conversations recognised the problem of finding a suitable fairness metric. Additionally, person C showed great awareness of the inherent ethical implications of deciding on which fairness metric to use and its effect on the treatment of different people by saying that there is a “complete ethical deliberation” behind picking a notion of fairness that requires one to think about how they “knowingly want to favour people” in order to achieve fairness as defined by that notion (C.2). This realisation supports the theoretical framework that was introduced in the introduction of this thesis and stresses the role fairness assessments play in finding a conception of fairness.

Fairness assessment tools

As pointed out in multiple conversations, there are tools available that are designed to guide the process of deciding on the most suitable fairness metric given the specific context. More specifically, the fairness tree is mentioned twice: both by person B and person D (B.5; D.4). The fairness tree is a decision tree for deciding which fairness metric to use in context of a specific algorithm. The fairness tree was designed as part of a bias and fairness audit (or assessment) toolkit named Aequitas (Saleiro et al., 2019). The fairness tree can be found in Appendix C: The Fairness Tree from the Aequitas Toolkit. An advantage of using the fairness tree can be that it enables data scientists to account for the use context of an algorithm in a concrete way that fits a technical way of thinking. As person D put it: “analysts (…) are glad when they see a fairness tree, because then they finally see something technical” (D.5). By being understandable for people with and without technical expertise, the fairness tree can also guide communication between data scientists and business stakeholders (as mentioned in excerpt B.5) and between data scientists and consultants (as mentioned in excerpt D.4).

However, the concreteness of the decision tree is also a downside, as it can lead to a narrow view of assessing fairness, reducing the complicated, normative task of deciding what should be considered (un)fair in context of a specific algorithm to a task of simply following a protocol by working through a decision tree. Regarding the choice of a suitable fairness metric as a protocol comes at the risk of a hyperfocus on the outcome of the protocol (the choice fairness metric and the numerical value that results from using it), requiring minimal understanding of the normative implications of this outcome. This concern is also noted by person D (D.5).

The concern that capturing fairness assessments in a standardised protocol will not provide the necessary understanding of the normative meaning of the outcome value of a fairness metric is not only relevant when using the fairness tree.[51] The more general problem appears to be that (even when conceptualizing fairness as non-discrimination) any method for assessing fairness that captures fairness into a numerical metric promotes a view of fairness that is too simplistic and ignores its normative meaning. Person C voiced a general mistrust towards tools that guide the process of algorithmic fairness assessments, because these tools often make assumptions about the definition of fairness[52], without making these assumptions and their consequences explicit. Hence person C states that these tools should only be used by one who understands how they work and what fairness notion they assume (C.3). Similar scepticism towards the use of fairness assessment tools because of their reliance on specific fairness metrics and the corresponding conceptions of fairness was voiced by one of the interviewees in the exploratory research by Costanza-Chock et al. (2022). The authors also found that 38% of their survey respondents did not use any such tools at all, which might also be explained by scepticism towards them.

Stakeholders

In the conversations that were held a few methods were mentioned that could help treading the ethical deliberation behind deciding on a fairness metric as well as the interpretation of the outcome of such

[51] In fact, one could argue that a merit of the fairness tree is that at least, it requires data scientists to consciously make decisions that result in a fairness metric, rather than simply picking a fairness metric because it is well known, or they know how to implement it.

[52] In the terminology introduced in this thesis, these tools can be said to implicitly decide on normative assessment choices, primarily the choice of how to test for fairness.

a metric with the proper care. As mentioned in excerpt B.5, the fairness tree was not used in isolation, but instead the fairness metric that was picked by using this tool was proposed to business stakeholders with more knowledge of the business context in which the algorithm was intended to be used. Consulting these stakeholders could help avoiding a view on fairness that is too technocentric. Person D also advocates the consultation of a diverse group in selecting a suitable fairness metric. However, they admit that it might be hard to explain the meaning of different fairness metrics to this group and that this might take too much time for people working at organisations using algorithmic decision-making (D.6). In their survey Costanza-Chock et al. (2022) found that 43% of the respondents who had answered the relevant question, indicated that stakeholder involvement is critical in assessing algorithmic fairness. However, they defined stakeholders as "those who are most at risk of harm" (Costanza-Chock et al., 2022, p. 1578), so the stakeholder group consulted by person B, which only consisted of people from within their organisation and no decision subjects, would not satisfy this definition. Similar to the remark in excerpt D.6 that explaining the meaning of fairness metrics to stakeholders is difficult, one of the interviewees of Costanza-Chock et al. also noted that in practice it is hard to get useful input from consulting stakeholders.

Given the importance of the explainability of fairness metrics to stakeholders, one could also use the degree of explainability as a selection criterium for a suitable fairness metric. This approach assumes that the communication and understanding of fairness metrics is more important than having a fairness metric that best captures the use context of an algorithm. This approach was chosen by person A and caused them to ultimately pick a variant of statistical parity (A.3). Person D also emphasises the importance of understandability of fairness metrics and proposes that this can be aided by expressing a fairness metric not just as a numerical value, but in a way that people are more likely to understand. E.g. instead of giving a disparate impact ratio, one could say one could that for each three women, ten men are hired (D.7).

In chapter 2 we saw that if statistical evidence is used in legal discrimination cases conceptually simple fairness metrics based on selection rates (such as statistical parity difference or disparate impact ratio) are also often selected for this. The purpose of this might also be to enable judges (who often lack technical expertise) to understand their meaning. These metrics might not always be most accurate within the use context of an algorithm, but as shown non-discrimination law has other ways to account for context.

Interpreting test results

What excerpt D.7 also shows and excerpt A.3 hints at, is that the use of calculating a fairness metric could be that it enables actors higher up in the organisation to judge whether they believe the use of an algorithm is acceptable given its measured fairness. This is in contrast with the suggestion from computer science of picking a fixed threshold value to decide whether a less then ideal fairness metric value is acceptable. It is also in contrast with using the use of the four-fifths rule in U.S. legal context. (See the chapter 1). Person B explained that an advice group of experts warned them not to use fixed thresholds, such as the four-fifths rule, because deciding whether a fairness metric value is acceptable is always a subjective decision. Person B believes this decision is a political choice that should be made by the person who has (political) responsibility over the decision-making process (B.6). Similarly, person D also believed that the main purpose of an algorithmic fairness assessment (consisting of documentation of all assessment choices and results) is informing the judgement of persons higher up in an organisation on whether an algorithm's fairness is acceptable (D.8).

Person A also recognised that the interpretation of the value of a fairness metric should involve people from people within the organisation using algorithmic decision-making, but outside of the team of data scientists doing a technical fairness assessment and they communicated the findings of their analysis to others within the organisation (A.4). Person A also believed that it is ultimately up to the managers of an organisation to determine what the organisation should strive for even if it is at an abstract level, so that subsequently the consequences of this moral ambition for the interpretation of a fairness metric value can be derived (A.5).

A common thread in excerpts B.6, D.8, A.4 and A.5 is that the ultimate judgement on whether the outcome of a fairness test is acceptable should not be with the team of data scientists who execute the test or based on fixed thresholds. Instead, it should be made by someone higher up in the organisation with (political) responsibility for the (moral) direction of the organisation. These excerpts clearly draw a line between what can be achieved by a team of data scientists (namely performing a technical fairness assessment) and where those higher up in the organisation should step in.

The assessment performed by person A remained without follow-up, possibly because those who should bear responsibility for the fairness of algorithms were not motivated to act upon these assessments (A.4; A.5). This problem does not appear to be isolated, as Costanza-Chock et al. (2022) found that 65% of their respondents reported that “auditees will not commit to address problems uncovered by audits” (p. 1577) (where they use a loose definition of audit, including many instances we would call assessments) and 80% reported that they had recommended at least one change to an AI system that was not implemented. Hence, it appears that the current lack of regulation that clearly mandates a committal audit framework introduces a great risk of algorithmic fairness assessments remaining without follow-up.

However, both person A and person B noted that testing fairness using metrics can also have a more straightforward use as well in comparing decision-making processes without the need to involve people higher up in the organisation. Both excerpt A.6 and B.7 show the value of calculating fairness metrics for the purpose of comparing the fairness of a proposed decision-making process relative to the fairness of the current process. The difference is that in A.6 a novel decision-making algorithm is compared to an established algorithm, whereas in B.7 the established decision-making method does not rely on an algorithm.

In conclusion, this section shows that a well performed and followed-up algorithmic fairness assessment requires good communication between data scientists and managers (or those with political responsibility). If the assessment is motivated by the personal interests of a group of data scientists, without management being inherently motivated to perform such an assessment, as was the case with the assessment performed by person A, there is a risk of the assessment remaining without proper follow-up. On the other hand, if the need for a fairness assessment is mostly felt by those higher up in an organisation there is a risk of data scientists perceiving the assessment as a purely technical exercise without understanding the moral consequences of their choices and findings. In choosing the suitable fairness metric for the context in which a specific algorithm is used, tools like the fairness tree can be helpful, although care should be taken that those using the tool do not blindly follow a protocol but understand and document the moral consequences of their choices and findings. From the perspective of understandability and the ability to communicate the meaning of the findings of a fairness assessment simple fairness metrics such as selection rates are preferable to more complicated ones.

Findings on data and re-evaluation

The fairness assessment methods discussed in chapter 1, require access to an evaluation data set that should be representative for the target population of the algorithm. However, in chapter 2 no clear requirements for the evaluation data were found and furthermore a trade-off was identified between the incentive to minimise data collection, as phrased in the EU *GDPR* and the storage of data ideally collected for assessing algorithmic fairness.

Although it is unclear what requirements evaluation data should have in order for fairness test that are performed on this data to be admissible as evidence in favour or against alleged algorithmic discrimination, it makes sense to demand at least that the population in this evaluation data should be representative of the target population of actual future decision subjects. This makes sense, because if this representativeness is not granted, statistics derived from this evaluation data do not provide any information on the actual impact the algorithm will have on demographic groups or the impact it can reasonably be expected to have.

Initial evaluation data selection

Unfortunately, the importance of the choice of a data set to evaluate fairness on was not identified yet by me when starting the conversations with practitioners. These conversations did cover the subject of data representativeness, but focused on whether the data used to *train* the model -instead of the data used to evaluate it- was representative for the subject population. Hence, when my conversations touched the subject of data representativeness it was often with the *bias in, bias out* principle in mind[53], and I wanted to know whether the practitioners performed tests to prevent that discriminatory bias would get *in* before testing whether it gets *out* using statistical fairness tests. It was only after having these conversations that my focus shifted to the question of whether the datafied reality captured by the evaluation data could be used to draw conclusions about the actual fairness of the algorithm in practice. However, if we assume that both the training data and fairness evaluation data are portions of the same data set that was gathered or created when developing the ML model, there is a connection between the representativeness of the training data and the representativeness of the fairness evaluation data. Hence the findings on representativeness of training data can still be relevant for the current scope.

In my conversations it turned out that the assumption that the population in the training data would be representative of the actual subject population and free of bias was often made without testing it. Both excerpts A.7 and B.8 show that historical data from the decision-making process was used to train (and assumingly evaluate) the ML model. In the case person B worked on, this was data from before an algorithm was involved in the decision-making process and in the case person A worked on this was data from a time in which an algorithm was already involved in the process. The population in this historical data is the population that has historically been the subject population for the decision-making process, unless there was a filter that selected only a part of the historical subject population to be stored and used in training and evaluating the model. (See chapter 1.) Excerpt B.8 explicitly mentions the absence of such filters. However, as excerpt A.7 shows, in the case person A worked on the training data consisted of placements, meaning that only those jobseekers who get selected for a job will end up in the training data. This means that the selection (or decision-making) process itself is a filter that selects only a specific part of the historical subject population (namely, those in the positive class) to be stored and used in training and evaluating the model. As discussed in chapter 1, since the algorithm was already involved in this decision-making process, there is a real risk that initial biases in the algorithm will be enlarged by this biased filtering mechanism. Excerpt A.8 shows that person A was aware of this risk and that their team attempted to alleviate it by complementing the training data with examples of decision subjects who were selected by humans instead of by the algorithm itself. This prevention of such a data feedback loop is also considered by Kearns & Roth (2019, Chapter 2). However, enriching the data in this way does not necessarily mean it will be representative, especially

[53] See chapter 1.

since excerpt A.9 shows that the data obtained from selection by humans is also filtered before it is added to the training data. Hence, the only way to really ensure the training data does not contain unacceptable, discriminatory bias is by testing for this. Unfortunately, the team of person A never got to this (A.10).

Keeping the *bias in, bias out* principle in mind, the absence of a test for discriminatory bias in the training data is worrisome. However, it should also be noted that the fairness metric ultimately used by person A and their team (a variant of demographic parity, see excerpt A.3) does not rely on the target labels in the evaluation data (as explained in chapter 1) and hence it is immune for traces of human bias within these target labels. Yet, on the other hand, since in this case human labelling also acts as filtering mechanism for the training (and presumably evaluation) data[54] human bias could still harm the representativeness of the evaluation data and thereby the validity of the fairness assessments on these data.[55]

In contrast to the aforementioned findings, one of the organisations which person D and their colleagues guided in their fairness assessment did test for the representativeness of training data for the historical subject population (D.9) They also tested whether the evaluation data was in turn representative of this training data (D.10). Costanza-Chock et al. (2022) found that 77% of the auditors responding to their survey reports they assess the quality of training data, although it is unclear whether and how such an assessment is concerned with representativeness for the subject population and bias against protected groups.

Monitoring and re-evaluation

Since it is impossible to predict the future and know exactly which persons with which features will be the future decision subjects, a fairness test always looks back into the past and is performed on historical data. Hence, at most such a test can tell whether an algorithm has or has not discriminated in the past (given that a fitting fairness metric was found, and the relevant protected groups are well-identified). This result is only informative for the future as long as the context in which the algorithm is employed does not change fundamentally. The most straightforward way of knowing whether an algorithm that was tested as non-discriminatory when it was deployed, was still non-discriminatory during a certain period in which it was in use, is by re-evaluating fairness on the decisions it aided during this period. Alternatively, one could also try to identify all contextual factors that need to be constant for a fairness assessment from the past to stay informative about the future and re-evaluate periodically whether these contextual factors remained constant indeed.

Person A, B and D all made plans for structural, periodical re-evaluation or monitoring of algorithmic fairness (A.11; B.9; D.11). However, for person A these plans never came into action because the plug was pulled for the algorithm altogether (A.11), for person B the plan did not yet come into action because the algorithm has not yet left the pilot stage (B.9) and as a consultant, person D was only involved in making plans for re-evaluation and not in performing it themselves (D.11).

[54] This is the case because only placements were included in the training data.

[55] E.g. if the decision-making process (both the algorithmic and human part) is biased against most women, but not against a certain subtype of atypical women (e.g. above average masculine women) most of the women that will end up in the training and evaluation data will be women of this atypical subtype. Then it might be the case that within this skewed evaluation data set no bias against women is measured since the algorithm is not biased against this atypical subtype of women who make up most of the women in the evaluation data. Yet, if the whole target population (with more typical women) was used for this bias analysis a bias against women would have been found.

Excerpts A.11, B.9 and D.11 all show an awareness of the importance of periodically reassessing fairness. Furthermore, excerpts B.9 and D.11 also show the use of a combination of quantitative and qualitative periodic tests or evaluations. Quantitative testing includes at least performing a bias analysis again as suggested above. Qualitative testing appears to be aimed at ensuring that important contextual factors in the use of the algorithm remained constant. This could include a legal compliance test and a (presumably qualitative) feature analysis. Excerpts B.9 and D.11 also show a strategy of formalizing the requirement for monitoring, either by making a specific person responsible for it (as suggested in excerpt B.9) or by incorporating it in existing protocols within an organisation (as suggested in excerpt D.11). Costanza-Chock et al. (2022) did not include questions about re-evaluation in their survey, although they do note that in (one-off) audits, quantitative methods are often preferred over qualitative ones. Furthermore, they show that 52% of their respondents reports they assess the "[e]xistence/quality of systems to report harm/appeal decisions" (p. 1575), which can be seen as a form of monitoring for harm.

In conclusion, the representativeness of the training (and presumably evaluation) data for the target population of an algorithm is not always tested, but sometimes rather assumed. Given that the evaluation data is a portion of a larger data set that is collected for training, validating and testing the algorithm, representativeness of the evaluation data for the target population of an algorithm requires that: (1) the evaluation data set is representative of this larger data set; (2) this larger data set is representative for the total historical target population of the decision process in which the algorithm is or will be applied and (3) this historical target population is representative for the current or future target population of the algorithm. Unfortunately, the practitioners in this research often simply assumed some or all these steps. However, all internal fairness assessors (including the ones person D provided consultation to) planned to periodically re-evaluate fairness. Assuming that this re-evaluation will be on data of all new decision subjects since the last evaluation (or a representative sample thereof), the first two requirements listed above would be met and although the third one can never be met with certainty, frequent re-evaluation arguably increases the chances that it will be met or at least that violation will be detected early on. Hence, the monitoring or re-evaluation plans of the practitioners are promising, but only worth something if they will be complied with in the future.

Key take-aways for the auditability of algorithmic fairness

Having presented the findings from our informal conversations, we will summarise the lessons that can be learnt from them and propose potential best practices that could be incorporated in algorithmic fairness audits. Of course, this identification is only meant as a starting point. Future (quantitative) research must show whether the practices suggested here are truly common and/or accepted among algorithmic fairness assessors and others who are or should be involved in this.

With respect to the question what protected attributes and groups to include in an algorithmic fairness assessment, there appears to be a division between "easy" and "hard" protected attributes to assess fairness for. Easy protected attributes, (e.g. sex and age) might already be stored by organisations and do not receive special protection from privacy regulation. Hard protected attributes (e.g. race or sexual orientation) are probably not stored already and storing them raises (legal) privacy concerns. Hence, a best practice might be to distinguish between these two types of protected attributes and require easy protected attributes to be tested for directly using suitable fairness metrics, while allowing hard protected attributes to be tested for indirectly. Methods for such indirect tests could include a qualitative input feature analysis and the calculation of fairness metrics for proxies (in a non-strict use of the word) of the attributes under consideration. The ability to assess intersectional discrimination is often hinder by a data availability problem. Methods of combatting this problem or methods of deciding when to test for intersectionality and when not, have not been identified in this chapter.

Several lessons can be learnt with respect to ways of testing for algorithmic fairness and interpreting the results of such tests. Firstly, the process of deciding on which test to use, can be aided by tools such as the decision tool or alternatively, by consulting (business) stakeholders. A good practice appears to be to combine both methods. In that case, the tools can aid communication between data scientists and people without technical expertise, while the involvement of people from outside a team of data scientists might prevent the use of the tool from becoming a purely technical procedure that ignores the ethical implications involved in choosing a fairness metric. In interpreting the outcome of a fairness metric two practices can be identified from our findings. Firstly, these outcomes can be communicated to people with (political) responsibility over the actions of an organisation to enable them to make a well-considered choice about whether to use an algorithm, given these metrics. In this case, conceptually simple metrics are preferred. Secondly, these outcomes can be used to compare several methods for a decision-making process and pick the one that is most fair (as measured by a certain metric).

Regarding the quality of the evaluation data used when performing a first algorithmic fairness assessment, a good practice could be to evaluate both the representativeness of the training data for the actual historical subject population and the representativeness of the evaluation data for the training data. Furthermore, in assuring that re-evaluations will happen frequently enough, good practices might be to designate responsibility for these re-evaluations to a specific person and/or to capture requirements for re-evaluation in the internal protocols of an organisation. Here, the tests included in re-evaluation could be both quantitative (e.g. calculating certain fairness metrics for new data) and qualitative (e.g. assessing whether the context in which the algorithm is used has changed in relevant ways.)

Conclusion and discussion

Conclusion

Algorithmic fairness can never be solely assessed from a computer science perspective, since in any attempt for such an assessment, normative assessment choices will inadvertently be faced. In this thesis we found that these choices centre around the definition of discrimination that should be used, the protected groups that should be considered, the way in which fairness should be tested and the data that should be used for this testing. If the goal of algorithmic fairness auditing is to prevent discrimination, the answer to the question of how to make these choices should be guided by non-discrimination legislation. For some choices (such as how to define discrimination or which protected attributes to use), this question is answered in Dutch non-discrimination legislation or at least the beginning of an answer can be found here. However, for other choices (such as what thresholds for fairness metrics to use or what the requirements for evaluation data to be a reliable source of statistical evidence) this question remains unanswered. Yet, in practice, the existence of normative assessment choices for which non-discrimination legislation does not provide guidance, does not necessarily have to prevent practitioners from performing algorithmic fairness assessments. Instead of guaranteeing non-discrimination, these assessments can have value as means of communication about algorithmic fairness within organisations. This favours conceptually simple fairness metrics such as statistical parity difference or disparate impact ratio. Another use of these assessments is in comparing different methods for the same decision-making process on a certain fairness metric. Furthermore, some normative assessment choices can be made by consulting (business) stakeholders.

Answering the question "What role can *auditing* play in ensuring algorithmic fairness, in terms of non-discrimination?" we conclude that external audits using objective and predefined audit criteria cannot be used to ensure non-discrimination, such that when passing this audit full compliance with non-discrimination legislation is ensured.

However, this does not mean auditing is useless in ensuring non-discrimination. Firstly, the internal algorithmic fairness assessments that are already performed might grow into first-party audits. This might be encouraged by demands of human rights conformity assessments in the upcoming EU *AI act*. This translation into audits would require the establishment of objective audit criteria. Of course, these criteria should not use strict thresholds for fairness metrics as this does not do justice to the contextuality of the meaning of fairness. Instead, these criteria should be of a more procedural nature. Instead of ensuring non-discrimination directly, the main goal of such an audit could be to ensure that (those with responsibility for the course of) an organisation was able to make a well-considered choice about whether to use an algorithm in light of possible discriminatory effects. An advantage of formalizing the requirements of a good internal algorithmic fairness assessment into an audit framework, is that it can be used across different organisations, so that not every organisation that wants to assess algorithmic fairness, must reinvent the wheel.

Best practices are a great source to base criteria for such a first-party algorithmic fairness audit. The findings of chapter 3 provide some idea of what those best practices could be. Of course, this chapter only provides a starting point for eventual audit criteria. Future quantitative research could confirm whether the practices identified here are commonly used and/or accepted among assessors of algorithmic fairness and others who are or should be involved in this. Once a list of best practices has been composed, their translation into strict audit criteria can also prove challenging.

It is highly questionable whether all organisations can be trusted to assess the fairness of their own algorithms or to audit this process. To ensure that all organisations (of a certain size) perform internal algorithmic fairness assessments and take them seriously, third-party audits can be used. These could either be executed by private audit firms or by legal oversight authorities. Again, these audits should have a procedural focus since the ultimate judgement on fairness cannot be captured in them. The criteria of these third-party audits could be like the criteria of the proposed first-party audits, although the third-party audits will be one step further removed from the actual algorithmic fairness assessment. Hence, they will more likely focus on the documentation of this assessment and can be used to ensure that an organisation has gathered and documented all information needed to judge about the fairness of its algorithmic decision-making processes. Furthermore, given the evaluation data, external auditors could check whether fairness metrics included in this documentation were properly calculated.

An advantage of mandating third-party audits is that it requires all organisations to assess algorithmic fairness, which might at least prevent some of the worst and most obvious cases of algorithmic discrimination. Furthermore, these audits can ensure that these organisations are all able to share the same set of information relevant to algorithmic fairness. This might aid the judiciary in judging over discrimination in the algorithmic domain. If policy makers truly want to enable decision subjects or advocacy organisations in establishing *prima facie* algorithmic discrimination more easily, they could also demand that certain key findings of algorithmic fairness assessments are always published. This could include the selection rates of certain protected groups, since this is often used as (basis for) statistical evidence in discrimination cases. Here, a third-party audit can ensure that these are the true selection rates of the algorithmic decision-making process as measured on evaluation data that meets certain criteria for representativeness of the actual population of decision subjects. (This would avoid organisations just making up fake selection rates that make their decision-making process appear non-discriminatory.) The conclusion that algorithmic fairness assessments could and should be used to empower the judiciary in ultimately judging on algorithmic discrimination is consistent with conclusions by other authors (Hildebrandt, 2020, Chapter 11.3; Wachter et al., 2020).

In conclusion, both internal and external auditing could play an important role in ensuring non-discrimination, albeit indirectly. One of the most pressing normative assessment choices that is not eliminated by law, namely how the outcome of fairness metrics or assessments should be interpreted, could be delegated. For the proposed form of first party audits, it is delegated to people with responsibility for the actions of an organisation and for the proposed form of third-party audits it is delegated to the judiciary. As Wachter et al. (2020) argue, the absence of threshold values in non-discrimination legislation is not a (unforeseen) shortcoming, but a (deliberate) strength, since it reflects that the meaning of discrimination is always dependent on contextual interpretation. Hence the delegation of the judgement on discrimination to humans that can interpret assessment results within the use context of an algorithm makes sense. Yet, a lot of choices still need to be made if one would desire an infrastructure in which first- and/or third-party audits help in ensuring algorithmic fairness. Some of the most pressing issues, which might need to be subject to public debate are discussed below.

Discussion

An interesting and important question raised by the findings of this thesis is whether and to what extent future regulations should eliminate all normative assessment choices that non-discrimination law currently does not eliminate. (E.g. Should law prescribe thresholds for fairness metrics?) On the one hand, legal scholars argue that contextual interpretation is key to law and therefore universal thresholds should not be given (Wachter et al., 2020). On the other hand, room for interpretation stands in the way of establishing audit frameworks for algorithmic fairness, which could provide organisations using algorithms in decision-making certainty about where they stand and whether they are susceptible to claims of algorithmic discrimination. Furthermore, these frameworks might enable algorithmic fairness to be assessed in a much more widespread and systematic way, than by having courts judge over individual cases.

Another problem highlighted by this thesis is the trade-off between privacy and the ability to assess algorithmic fairness. All metrics for algorithmic (group) fairness rely on the availability of data on protected attributes. However, many features that are considered so sensitive that they should be protected against discrimination are also considered so sensitive that, in principle, data on them should not be gathered or stored to respect the privacy of data subjects. In practice this can lead to organisations only testing for potential discrimination of protected groups defined by protected attributes which are relatively uncontroversial or legally permitted to gather, such as sex or age. This means that discrimination against groups defined by more privacy sensitive protected attributes, such as sexual orientation, religion, political opinion, disability or race might remain completely unnoticed by fairness assessments. Here the important question is whether the importance of assessing fairness is greater than the importance of safeguarding privacy. Yet, it should be noticed that technological advancements might circumvent this trade-off by enabling fairness assessments without needing to store protected attributes that can be traced to individual decision subjects.

Lastly, it is important to compare algorithmic decision-making to human decision-making in terms of risks of unfairness and opportunities to ensure fairness. The examples of algorithmic discrimination given in the introduction of and throughout this thesis might have left the reader with the impression that algorithmic decision-making is a dangerous endeavour that constitutes a high risk of discrimination. Judging from conversations I have in my personal life news items about these cases of algorithmic discrimination seem to have made many people suspicious of algorithmic decision-making indeed. This is not completely unjustified, as algorithmic decision-making introduces new risks of discrimination at a large scale that can remain unnoticed due to an overreliance on automation and cannot always be handled well by a legal system that was not designed to deal with algorithms.

However, one should not forget that human decision-making often leads to terrible discrimination as well. In fact, algorithmic discrimination is partly a replication of historic human discrimination.[56] It is senseless to ask whether algorithmic decision-making is more prone to discrimination than human decision-making or the other way around, since this completely depends on how the decision-making process is designed and what measures are taken to prevent discrimination. Yet, I do want to note that besides raising new issues, algorithmic decision-making also gives rise to many new opportunities in combatting discrimination. Because of the emergence of algorithms in decision-making processes, much more data about decision outcomes is generated and stored. This has led to an abundant amount of algorithmic fairness literature on how this data can be used to get insight into the fairness of algorithms, enabling much better detection of potential discrimination. Although the issues raised by (improper use of) algorithmic decision-making are most likely a very important reason why so many scandals of algorithmic discrimination have become public, another reason for this could be that the availability of this data on decision outcomes enabled (journalistic) investigators to make a strong case for discrimination. A human decision-making process might have discriminated just as much, but we might have never found out because no data was stored to prove this.[57] Hence, if the relevant data is

[56] Historical human discrimination can cause training data to be biased against the discriminated group, which means that (without proper mitigation), ML models trained on this data will replicate this discriminatory bias.

[57] Of course, it is possible to keep data of a fully human decision-making process as well by requiring every decision maker to record all decisions they make, but in practice storage of decision data is much more common for algorithms, since this data is needed to evaluate their performance over time anyway. Furthermore, this data consists not only of decision outcomes but also of a strictly defined set of input features. In many cases of human

made publicly available, in some cases algorithmic discrimination might be *less* instead of more likely to go unnoticed than human discrimination. Additionally, another part of algorithmic fairness literature proposes many possibilities to reduce the risk of algorithmic discrimination.

In the end, algorithmic decision-making seems primarily dangerous when used without caution. Although the possibilities for detecting and preventing algorithmic discrimination are great, it is also very easy not to do this, not to test for fairness in anyway, not to share any data on the decision outcomes of the algorithm or even disclose that it is used in the first place. Many scandals of algorithmic discrimination seem to be caused by algorithms being implemented and never cared for again. Once more, this strengthens the need for binding forms of algorithmic fairness regulation. Auditing can play an important role.

Limitations

The most important limitations of my research have already been discussed throughout this thesis. Firstly, the meaning of *algorithmic fairness* within this thesis was limited to non-discrimination, excluding a wide range of algorithmic harms that could also be called unfair. Secondly, as an AI student my understanding of law and sociotechnical systems is way below the level of legal and social science students, respectively. One could argue that the only way to truly respect the multidisciplinary of algorithmic fairness would be to have a multidisciplinary team researching it but given the limitations of a master thesis project I settled with an investigation into the multidisciplinary aspects of algorithmic fairness from the perspective of a (broadly interested) computer scientist. Thirdly, the informal conversations supporting chapter 3 of this thesis were held before all findings from the two preceding chapters were fully solidified, meaning that they contained little information about a few normative assessment choices. Notwithstanding these issues I think this research has value as a computer scientific contribution to the interdisciplinary discourse on algorithmic fairness.

decision-making, it might not be clear what factors (features) influenced a decision, let alone that they would be stored.

Literature

Note on referencing system: This thesis uses the APA citation style (7th edition). However, since this style is less suited for referencing legal sources, such as laws, directives, regulations, court rulings, etc. since these often do not have a clear author, and a long official name, those are referenced in-text by using an abbreviation of their official name. These abbreviations are always in Italics and the information needed to retrieve these sources is contained in a separate list that follows the regular APA-style bibliography.

APA-style references

ACM FAccT. (n.d.). *ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*. Retrieved February 8, 2024, from https://facctconference.org/

Alao, R., Bogen, M., Miao, J., Mironov, I., & Tannen, J. (2021). *How Meta is working to assess fairness in relation to race in the U.S. across its products and systems*. https://ai.meta.com/research/publications/how-meta-is-working-to-assess-fairness-in-relation-to-race-in-the-us-across-its-products-and-systems/

Alexander, L. (1992). What Makes Wrongful Discrimination Wrong? Biases, Preferences, Stereotypes, and Proxies. *University of Pennsylvania Law Review*, *141*(1), 149. https://doi.org/10.2307/3312397

Alexander, L., & Cole, K. (1997). Discrimination by Proxy. *Constitutional Commentary*, *14*. https://heinonline.org/HOL/Page?handle=hein.journals/ccum14&id=461&div=29&collection=journals

Algemene Rekenkamer. (2020). *Digitaal Toetsingskader Algoritmes*. https://www.rekenkamer.nl/publicaties/publicaties/2021/01/28/download-het-toetsingskader

Algemene Rekenkamer. (2022). *Algoritmes getoetst: De inzet van 9 algoritmes bij de overheid*. https://www.rekenkamer.nl/publicaties/rapporten/2022/05/18/algoritmes-getoetst

Amaro, S. (2021, January 15). *Dutch government resigns after childcare benefits scandal*. CNBC. https://www.cnbc.com/2021/01/15/dutch-government-resigns-after-childcare-benefits-scandal-.html

Angwin, J., Larson, J., Mattu, S., Kirchner, L., & ProPublica. (2016, May 23). Machine Bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Autoriteit Persoonsgegevens. (2020). *Belastingdienst/Toeslagen - De verwerking van de nationaliteit van aanvragers van kinderopvangtoeslag*. https://www.autoriteitpersoonsgegevens.nl/documenten/onderzoek-belastingdienst-kinderopvangtoeslag

Avbelj, M. (2011). Supremacy or Primacy of EU Law—(Why) Does it Matter? *European Law Journal*, *17*(6), 744–763. https://doi.org/10.1111/J.1468-0386.2011.00560.X

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and Opportunities*. MIT Press. https://fairmlbook.org/

Barta, G. (2018). The increasing role of IT auditors in financial audit: risks and intelligent answers. *Business, Management and Education*, *16*(1), 81–93. https://doi.org/10.3846/bme.2018.2142

Bell, J. (1997). Book Review Comparing Precedent. *Cornell Law Review*, *82*, 1243–1278. https://heinonline.org/HOL/Page?handle=hein.journals/clqv82&id=1299&div=53&collection=journals

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *IBM Journal of Research and Development*, *63*(4/5), 4:1-4:15. https://arxiv.org/abs/1810.01943v1

Belleman, B., Heilbron, B., & Kootstra, A. (2023, June 21). *De discriminerende fraudecontroles van Duo*. Investico. https://www.platform-investico.nl/onderzoeken/de-discriminerende-fraudecontroles-van-duo

Belli, L., Yee, K., Tantipongpipat, U., Gonzales, A., Lum, K., & Hardt, M. (2023). *County-level Algorithmic Audit of Racial Bias in Twitter's Home Timeline*. https://doi.org/10.48550/arXiv.2211.08667

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? 🦜. *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017). Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods and Research*, *50*(1), 3–44. https://doi.org/10.1177/0049124118782533

Besse, P., del Barrio, E., Gordaliza, P., Loubes, J. M., & Risser, L. (2022). A Survey of Bias in Machine Learning Through the Prism of Statistical Parity. *The American Statistician*, *76*(2), 188–198. https://doi.org/10.1080/00031305.2021.1952897

Bietti, E. (2020). From ethics washing to ethics bashing: A view on tech ethics from within moral philosophy. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 210–219. https://doi.org/10.1145/3351095.3372860

Binns, R. (2020). On the apparent conflict between individual and group fairness. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 514–524. https://doi.org/10.1145/3351095.3372864

Blakeney, C., Atkinson, G., Huish, N., Yan, Y., Metsis, V., & Zong, Z. (2022). Measuring Bias and Fairness in Multiclass Classification. *2022 IEEE International Conference on Networking, Architecture and Storage (NAS)*, 1–6. https://doi.org/10.1109/NAS55553.2022.9925287

Brown, S., Davidovic, J., & Hasan, A. (2021). The algorithm audit: Scoring the algorithms that score us. *Big Data and Society*, *8*(1). https://doi.org/10.1177/2053951720983865

Bullock, J., & Masselot, A. (2012). Multiple Discrimination and Intersectional Disadvantages: Challenges and Opportunities in the European Union Legal Framework. *Columbia Journal of European Law*, *19*. https://heinonline.org/HOL/Page?handle=hein.journals/coljeul19&id=63&div=6&collection=journals

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Vol. 81, pp. 77–91). PMLR. https://proceedings.mlr.press/v81/buolamwini18a.html

Cabrera, A. A., Epperson, W., Hohman, F., Kahng, M., Morgenstern, J., & Chau, D. H. (2019). FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning. *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 46–56. https://doi.org/10.1109/VAST47406.2019.8986948

Cambridge Dictionary. (n.d.). *algorithm*. Retrieved July 10, 2023, from https://dictionary.cambridge.org/dictionary/english/algorithm

Campbell, C., & Smith, D. (2023). Distinguishing Between Direct and Indirect Discrimination. *The Modern Law Review*, *86*(2), 307–330. https://doi.org/10.1111/1468-2230.12760

Canbek, G. (2022). Gaining insights in datasets in the shade of "garbage in, garbage out" rationale: Feature space distribution fitting. *WIREs Data Mining and Knowledge Discovery*, *12*(3), e1456. https://doi.org/https://doi.org/10.1002/widm.1456

Carrier, R. (2021). Infrastructure of Trust for AI - Guide to Entity Roles and Responsibilities. In *ForHumanity*. https://forhumanity.center/article/infrastructure-of-trust-for-ai-guide-to-entity-roles-and-responsibilities/

Carrier, R., & Brown, S. (2021). Taxonomy: AI Audit, Assurance & Assessment. In *ForHumanity*. https://forhumanity.center/article/taxonomy-ai-audit-assurance-assessment/

Castelnovo, A., Crupi, R., Gamba, G. Del, Greco, G., Naseer, A., Regoli, D., & Miguel Gonzalez, B. S. (2020). BeFair: Addressing Fairness in the Banking Sector. *Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020*, 3652–3661. https://doi.org/10.1109/BIGDATA50022.2020.9377894

Caton, S., & Haas, C. (2023). Fairness in Machine Learning: A Survey. *ACM Comput. Surv.* https://doi.org/10.1145/3616865

Chen, J., Kallus, N., Mao, X., Svacha, G., & Udell, M. (2019). Fairness under unawareness: Assessing disparity when protected class is unobserved. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 339–348. https://doi.org/10.1145/3287560.3287594

Chouldechova, A. (2017). Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163. http://dx.doi.org/10.1089/big.2016.0047

Claes, M. (2015). The Primacy of EU Law in European and National Law. In *The Oxford Handbook of European Union Law*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199672646.013.8

College voor de Rechten van de Mens. (2020, May 2). *Rechter oordeelt dat fraudedetectiesysteem SyRI in strijd is met mensenrechten*. https://www.mensenrechten.nl/actueel/toegelicht/toegelicht/2020/rechter-oordeelt-dat-fraudedetectiesysteem-syri-in-strijd-is-met-mensenrechten

College voor de Rechten van de Mens. (2022a). *Monitor Discriminatiezaken 2022*. https://publicaties.mensenrechten.nl/publicatie/fa21a1f1-4efd-402e-a868-784b45e2b93f

College voor de Rechten van de Mens. (2022b). *Vooronderzoek naar de vermeende discriminerende effecten van de werkwijzen van de Belastingdienst/Toeslagen*. https://publicaties.mensenrechten.nl/publicatie/a356efac-8752-4843-a86c-665bb0a98667

Cooper, R., & Foster, M. (1971). Sociotechnical systems. *American Psychologist*, *26*(5), 467–474. https://doi.org/10.1037/h0031539

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. https://doi.org/10.1145/3097983.3098095

Cornacchia, G., Anelli, V. W., Biancofiore, G. M., Narducci, F., Pomo, C., Ragone, A., & Di Sciascio, E. (2023). Auditing fairness under unawareness through counterfactual reasoning. *Information Processing & Management*, *60*(2), 103224. https://doi.org/10.1016/J.IPM.2022.103224

Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022). Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. *ACM International Conference Proceeding Series*, 1571–1583. https://doi.org/10.1145/3531146.3533213

Coston, A., Ramamurthy, K. N., Wei, D., Varshney, K. R., Speakman, S., Mustahsan, Z., & Chakraborty, S. (2019). Fair transfer learning with missing protected attributes. *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 91–98. https://doi.org/10.1145/3306618.3314236

Council of Europe, European Court of Human Rights, & European Union Agency for Fundamental Rights. (2018). *Handbook on European non-discrimination law – 2018 edition* (2018 edition). Publications Office of the European Union. https://doi.org/doi/10.2811/792676

Council of the EU. (2023, December 9). *Artificial intelligence act: Council and Parliament strike a deal on the first rules for AI in the world - Consilium*. Council of the EU. https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/

Crampton, N. (2022, June 21). *Microsoft's framework for building AI systems responsibly*. https://blogs.microsoft.com/on-the-issues/2022/06/21/microsofts-framework-for-building-ai-systems-responsibly/

Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press. https://doi.org/10.1007/s00146-022-01488-x

Croak, M. (2023, January 23). *Google Research, 2022 & beyond: Responsible AI*. https://ai.googleblog.com/2023/01/google-research-2022-beyond-responsible.html

Cumbo, L. A., Ampry-Samuel, A., Rosenthal, H. K., Cornegy, R. E., Kallos, B., Adams, A. E., Louis, F. N., Chin, M. S., Cabrera, F., Rose, D. L., Gibson, V. L., Brannan, J. L., Rivera, C., Levine, M., Ayala, D. I., Miller, I. D., Levin, S. T., & Barron, I. D. (2021). *A Local Law to amend the administrative code of the city of New York, in relation to automated employment decision tools* (1894–2020). The New York City Council. https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&Options=Advanced&Search

Dastin, J. (2018, October 10). *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Dastin, J., & Paresh, D. (2021, February 4). *Two Google engineers resign over firing of AI ethics researcher Timnit Gebru*. Reuters. https://www.reuters.com/article/us-alphabet-resignations/two-google-engineers-resign-over-firing-of-ai-ethics-researcher-timnit-gebru-idUSKBN2A4090

Dolata, M., Feuerriegel, S., & Schwabe, G. (2022). A sociotechnical view of algorithmic fairness. *Information Systems Journal*, *32*(4), 754–818. https://doi.org/10.1111/ISJ.12370

Draude, C., Klumbyte, G., Lücking, P., & Treusch, P. (2020). Situated algorithms: a sociotechnical systemic approach to bias. *Online Information Review*, *44*(2), 325–342. https://doi.org/10.1108/OIR-10-2018-0332/FULL/PDF

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *ITCS 2012 - Innovations in Theoretical Computer Science Conference*, 214–226. https://doi.org/10.1145/2090236.2090255

Dworkin, R. (1972). The Jurisprudence of Richard Nixon. *The New York Review of Books*, *18*(8), 27–35. https://www.nybooks.com/articles/1972/05/04/a-special-supplement-the-jurisprudence-of-richard/

Ersoy, S., & van der Gaag, S. (2023, June 21). *Studenten met migratieachtergrond opvallend vaak beschuldigd van fraude, minister wil systeem grondig nagaan*. NOS Op 3. https://nos.nl/op3/artikel/2479700-studenten-met-migratieachtergrond-opvallend-vaak-beschuldigd-van-fraude-minister-wil-systeem-grondig-nagaan

Escalante, H. J., Kaya, H., Salah, A. A., Escalera, S., Güçlütürk, Y., Güçlü, U., Baró, X., Guyon, I., Junior, J. C. S. J., Madadi, M., Ayache, S., Viegas, E., Gürpınar, F., Wicaksana, A. S., Liem, C. C. S., Gerven, M. A. J. van, & Lier, R. van. (2022). Modeling, Recognizing, and Explaining Apparent Personality From Videos. *IEEE Transactions on Affective Computing*, *13*(2), 894–911. https://doi.org/10.1109/TAFFC.2020.2973984

European Parliament. (2024, February 13). *Artificial Intelligence Act: committees confirm landmark agreement*. https://www.europarl.europa.eu/news/en/press-room/20240212IPR17618/artificial-intelligence-act-committees-confirm-landmark-agreement

European Union. (n.d.). *Court of Justice of the European Union*. Retrieved February 28, 2024, from https://european-union.europa.eu/institutions-law-budget/institutions-and-bodies/search-all-eu-institutions-and-bodies/court-justice-european-union-cjeu_en

Facebook's Civil Rights Audit-Final Report. (2020). https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and Removing Disparate Impact. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268. https://doi.org/10.1145/2783258.2783311

Floridi, L. (2019). Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. *Philosophy and Technology*, *32*(2), 185–193. https://doi.org/10.1007/S13347-019-00354-X

Fox, W. M. (1995). Sociotechnical System Principles and Guidelines: Past and Present. *Https://Doi.Org/10.1177/0021886395311009*, *31*(1), 91–105. https://doi.org/10.1177/0021886395311009

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*, *23*(3), 551–567. https://doi.org/10.1007/S10676-020-09577-5

Freudenburg, W. R. (2003). Social Impact Assessment. *Https://Doi.Org/10.1146/Annurev.so.12.080186.002315*, *12*(1), 451–478. https://doi.org/10.1146/ANNUREV.SO.12.080186.002315

Galhotra, S., Shanmugam, K., Sattigeri, P., & Varshney, K. R. (2021). Interventional Fairness with Indirect Knowledge of Unobserved Protected Attributes. *Entropy*, *23*(12). https://doi.org/10.3390/e23121571

Gallie, W. B. (1955). Essentially Contested Concepts. *Proceedings of the Aristotelian Society*, *56*, 167–198. http://www.jstor.org/stable/4544562

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92. https://doi.org/10.1145/3458723

Geiger, G. (2021, March 1). *How a Discriminatory Algorithm Wrongly Accused Thousands of Families of Fraud*. https://www.vice.com/en/article/jgq35d/how-a-discriminatory-algorithm-wrongly-accused-thousands-of-families-of-fraud

Geiger, G., Constantaras, E., Braun, J.-C., Aung, H., Schot, E., Klassen, S., Van Dijk, R., Davidson, D., Mehrota, D., Meaker, M., Burgess, M., Thomas, K., Walker, A., Lam, K., Lavigne, S., Qu, A., Nagendran, R., Moorthy, H., Tiwari, I., … Kollias, F. (2023, March 2). Suspicion Machines. *Lighthouse Reports*. https://www.lighthousereports.com/investigation/suspicion-machines/

Gerards, J., Schäfer, M. T., Muis, I., & Vankan, A. (2022). *Fundamental Rights and Algorithms Impact Assessment (FRAIA)*. https://dspace.library.uu.nl/handle/1874/420552

Goodman, B. W. (2016). A step towards accountable algorithms?: Algorithmic discrimination and the European Union general data protection. *9th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona. NIPS Foundation*. http://www.mlandthelaw.org/papers/goodman1.pdf

Goodman, E. P., & Trehu, J. (2022). AI Audit Washing and Accountability. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.4227350

Greenberg, I. (1979). An Analysis of the EEOCC “Four-Fifths” Rule. *Source: Management Science*, *25*(8), 762–769. https://www.jstor.org/stable/2630312?seq=1&cid=pdf-

Hacker, P. (2018). Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law. *Common Market Law Review*, 1143–1185. http://www.kluwerlawonline.com/api/Product/CitationPDFURL?file=Journals\COLA\COLA2018095.pdf

Hajian, S., & Domingo-Ferrer, J. (2013). A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering*, *25*(7), 1445–1459. https://doi.org/10.1109/TKDE.2012.72

Hardt, M., Price, E., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems*, *29*. https://doi.org/10.48550/arXiv.1610.02413

Hellman, D. (2020). MEASURING ALGORITHMIC FAIRNESS. *Virginia Law Review*, *106*(4), 811–866. https://www.jstor.org/stable/27074708

Hellström, T., Dignum, V., & Bensch, S. (2020). Bias in Machine Learning -- What is it Good for? *CEUR Workshop Proceedings*, *2659*, 3–10. https://arxiv.org/abs/2004.00686v2

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

Herlitz, A. (2022). Predictive Fairness. *Philosophical Studies Series*, *143*, 141–161. https://doi.org/10.1007/978-3-030-75267-5_5/COVER

Hertweck, C., Heitz, C., & Loi, M. (2021). On the moral justification of statistical parity. *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 747–757. https://doi.org/10.1145/3442188.3445936

Hildebrandt, M. (2020). Law for Computer Scientists and Other Folk. In *Law for Computer Scientists and Other Folk* (online). Oxford Academic. https://doi.org/10.1093/OSO/9780198860877.001.0001

Hyde, S. J., Bachura, E., & Harrison, J. S. (2023). Garbage in, Garbage out: A Theory-Driven Approach to Improve Data Handling in Supervised Machine Learning. In A. D. Hill, A. F. McKenny, P. O'Kane, & S. Paroutis (Eds.), *Methods to Improve Our Field* (Vol. 14, pp. 101–132). Emerald Publishing Limited. https://doi.org/10.1108/S1479-838720220000014006

Iliadis, A., & Russo, F. (2016). Critical data studies: An introduction. *Http://Dx.Doi.Org/10.1177/2053951716674238*, *3*(2). https://doi.org/10.1177/2053951716674238

International Organization for Standardization. (2018). *Guidelines for auditing management system (ISO Standard No. 19011:2018)*. https://www.iso.org/obp/ui/#iso:std:iso:19011:ed-3:v1:en

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R* (2nd ed.). Springer Science+Business Media, LLC. http://www.springer.com/series/417

Johnson, K. R., & Martinez, G. A. (1999). Discrimination by Proxy: The Case of Proposition 227 and the Ban on Bilingual Education. *U.C. Davis Law Review*, *33*. https://heinonline.org/HOL/Page?handle=hein.journals/davlr33&id=1239&div=43&collection=journals

Kak, A., & Myers West, S. (2023). *AI Now 2023 Landscape: Confronting Tech Power*. https://ainowinstitute.org/2023-landscape.

Kallus, N., Mao, X., & Zhou, A. (2021). Assessing Algorithmic Fairness with Unobserved Protected Class Using Data Combination. *Management Science*, *68*(3), 1959–1981. https://doi.org/10.1287/mnsc.2020.3850

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33. https://doi.org/10.1007/S10115-011-0463-8/METRICS

Kearns, M., & Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press. https://dl.acm.org/doi/book/10.5555/3379082

Kim, J. Y., Ortiz, C., Nam, S., Santiago, S., & Datta, V. (2020). *Intersectional Bias in Hate Speech and Abusive Language Datasets*. https://arxiv.org/abs/2005.05921v3

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

Kusner, M., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual Fairness. *Advances in Neural Information Processing Systems*, *2017-December*, 4067–4077. https://arxiv.org/abs/1703.06856v3

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 275–284. https://doi.org/10.1145/3097983.3098066

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016, May 23). *How We Analyzed the COMPAS Recidivism Algorithm*. ProPublica. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Laufer, B., Jain, S., Cooper, A. F., Kleinberg, J., & Heidari, H. (2022). Four Years of FAccT: A Reflexive, Mixed-Methods Analysis of Research Contributions, Shortcomings, and Future Prospects. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 401–426. https://doi.org/10.1145/3531146.3533107

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature 2015 521:7553*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

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 and Technology*, *31*(4), 611–627. https://doi.org/10.1007/S13347-017-0279-X/METRICS

Loof, J. P. (2020). Toekomstige uitdagingen voor het gelijkebehandelingsrecht. *NTM|NJCM-Bull*, *45*(2), 251–277. https://hdl.handle.net/1887/3200798

Lucaj, L., van der Smagt, P., & Benbouzid, D. (2023). AI Regulation Is (not) All You Need. *2023 ACM Conference on Fairness, Accountability, and Transparency*, *13*(23), 1267–1279. https://doi.org/10.1145/3593013.3594079

Makhlouf, K., Zhioua, S., & Palamidessi, C. (2021). On the Applicability of Machine Learning Fairness Notions. *ACM SIGKDD Explorations Newsletter*, *23*(1), 14–23. https://doi.org/10.1145/3468507.3468511

Makkonen, T. (2007). *Measuring Discrimination: Data Collection and EU Equality Law*. http://hdl.handle.net/20.500.12389/19825

Maliszewska-Nienartowicz, J. (n.d.). Direct and Indirect Discrimination in European Union Law-How to Draw a Dividing Line? ∗. *International Journal of Social Sciences*, *III*(1), 2014. Retrieved November 23, 2023, from https://api.semanticscholar.org/CorpusID:53619830

Mayson, S. G. (2019). Bias in, Bias out. *Yale Law Journal*, *128*(8), 2218–2301. https://heinonline.org/HOL/P?h=hein.journals/ylr128&i=2309

Md Ali, A., & Teck-Heang, L. (2008). The evolution of auditing: An analysis of the historical development. *Journal of Modern Accounting and Auditing*, *4*(12), 1548–6583. https://www.researchgate.net/publication/339251518

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys (CSUR)*, *54*(6). https://doi.org/10.1145/3457607

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. https://doi.org/10.1145/3287560.3287596

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. https://doi.org/10.1146/annurev-statistics-042720

Mökander, J., Axente, M., Casolari, F., & Floridi, L. (2022). Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation. *Minds and Machines*, *32*(2), 241–268. https://doi.org/10.1007/S11023-021-09577-4

Mökander, J., & Floridi, L. (2021). Ethics-based auditing to develop trustworthy AI. In *Minds and Machines*. doi.org/10.1007/s11023-021-09557-8

Mukherjee, D., Yurochkin, M., Banerjee, M., & Sun, Y. (2020). Two Simple Ways to Learn Individual Fairness Metrics from Data. *Proceedings of the 37th International Conference on Machine Learning*. https://proceedings.mlr.press/v119/mukherjee20a.html.

Nachbar, T. B. (2021). Algorithmic Fairness, Algorithmic Discrimination. *Florida State University Law Review*, *48*(2), 509–558. https://heinonline.org/HOL/Page?handle=hein.journals/flsulr48&id=533&div=15&collection=journals

NOS Nieuws. (2021, January 15). *Kabinet-Rutte III gevallen; Wiebes helemaal weg*. NOS Nieuws. https://nos.nl/collectie/13855/artikel/2364513-kabinet-rutte-iii-gevallen-wiebes-helemaal-weg

NOS Nieuws. (2024, January 22). *Bijna 70.000 mensen meldden zich voor deadline als "toeslagenouders."* NOS Nieuws. https://nos.nl/artikel/2505816-bijna-70-000-mensen-meldden-zich-voor-deadline-als-toeslagenouders

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science* , *366*(6464), 447–453. https://doi.org/10.1126/SCIENCE.AAX2342

ORCAA. (2020). *Description of Algorithmic Audit: Pre-built Assessments*. https://www.hirevue.com/resources/template/orcaa-report

Pesenti, J. (2021, June 22). *Facebook's five pillars of Responsible AI*. Meta AI Blog. https://ai.meta.com/blog/facebooks-five-pillars-of-responsible-ai/

Pessach, D., & Shmueli, E. (2023). Algorithmic Fairness. In L. Rokach, O. Maimon, & E. Shmueli (Eds.), *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook* (pp. 867–886). Springer International Publishing. https://doi.org/10.1007/978-3-031-24628-9_37

Prince, A. E. R., & Schwarcz, D. (2019). Proxy Discrimination in the Age of Artificial Intelligence and Big Data. *Iowa Law Review*, *105*. https://heinonline.org/HOL/Page?handle=hein.journals/ilr105&id=1283&div=35&collection=journals

Radiya-Dixit, E., & Neff, G. (2023). A Sociotechnical Audit: Assessing Police Use of Facial Recognition. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 1334–1346. https://doi.org/10.1145/3593013.3594084

Rai, N. (2021). Why ethical audit matters in artificial intelligence? *AI and Ethics*, *2*(1), 209–218. https://doi.org/10.1007/S43681-021-00100-0

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. https://doi.org/10.1145/3351095.3372873

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., & Goldberg, Y. (2020). Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 7237–7256. https://doi.org/10.18653/v1/2020.acl-main.647

Rekenkamer Rotterdam. (2021, April 15). *gekleurde technologie - verkenning ethisch gebruik algoritmes*. https://rekenkamer.rotterdam.nl/onderzoeken/algoritmes/

Romanov, A., De-Arteaga, M., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., Rumshisky, A., & Kalai, A. T. (2019). What's in a Name? Reducing Bias in Bios without Access to Protected Attributes. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies - Proceedings of the Conference, *1*, 4187–4195. https://arxiv.org/abs/1904.05233v1

Ropohl, G. (1999). Philosophy of Socio-Technical Systems. *Society for Philosophy and Technology Quarterly Electronic Journal*, *4*(3), 186–194. https://doi.org/10.5840/TECHNE19994311

Russell, S. J., & Norvig, P. (2010). *Artificial Intelligence A Modern Approach* (3rd ed.). Pearson Education, Inc. http://aima.cs.berkeley.edu/

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T., & Ghani, R. (2019). *Aequitas: A Bias and Fairness Audit Toolkit*. https://arxiv.org/abs/1811.05577v2

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*. https://websites.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf

Schäfer, M. T., Van Es, K., & Muis, I. (2023). Investigating the Datafied Society. In K. van Es & N. Verhoeff (Eds.), *Situating Data Inquiries in Algorithmic* (pp. 267–272). Amsterdam University Press. https://dspace.library.uu.nl/handle/1874/420552

Schiek, D., & Lawson, A. (Eds.). (2016). *European Union non-discrimination law and intersectionality : investigating the triangle of racial, gender and disability discrimination*. Routledge. https://www.routledge.com/European-Union-Non-Discrimination-Law-and-Intersectionality-Investigating/Schiek-Lawson/p/book/9781138269453

Sephus, N. (2022, October 19). *Responsible Use of Artificial Intelligence and Machine Learning [recorded presentation]*. https://pages.awscloud.com/Responsible-Use-of-Artificial-Intelligence-and-Machine-Learning_2022_1007-MCL_OD

Shaughnessy, B., Braun, S., Hentschel, T., & Peus, C. V. (2016). Diverse and just? The role of quota-based selection policies on organizational outcomes. *European Journal of Social Psychology*, *46*(7), 880–890. https://doi.org/10.1002/EJSP.2208

Shelby, R., Rismani, S., Henne, K., Moon, Aj., Rostamzadeh, N., Nicholas, P., Yilla-Akbari, N., Gallegos, J., Smart, A., Garcia, E., & Virk, G. (2023). Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 723–741. https://doi.org/10.1145/3600211.3604673

Shin, D. D. H. (2019). *Socio-Technical Design of Algorithms: Fairness, Accountability, and Transparency*. http://hdl.handle.net/10419/205212

Simonite, T. (2021, May 8). What Really Happened When Google Ousted Timnit Gebru. *Wired*. https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened/

Spielkamp, M. (2023, January 19). *New project: Auditing Algorithms for Systemic Risks*. AlgorithmWatch. https://algorithmwatch.org/en/new-project-auditing-algorithms-for-systemic-risks/

Steinberg, D., Reid, A., & O'Callaghan, S. (2020). *Fairness Measures for Regression via Probabilistic Classification*. https://arxiv.org/abs/2001.06089v2

Straatman, J., Muis, I., Van Der Weijden, D., & Schäfer, M. T. (2023). Beyond the Algorithm Audit: Developing a Recurrent Method for Monitoring Algorithmic Systems from a Socio-Technical Perspective. In *[unpublished research paper]*.

Stuart Geiger, R., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., & Huang, J. (2020). Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 325–336. https://doi.org/10.1145/3351095.3372862

The Supreme Audit Institution of Finland, The Supreme Audit Institution of Germany, The Supreme Audit Institution of the Netherlands, The Supreme Audit Institution of Norway, & The Supreme Audit Institution of the UK. (2023). *Auditing machine learning algorithms: A white paper for public auditors*. https://auditingalgorithms.net/

Theodoridis, S., & Koutroumbas, K. (2006). *Pattern Recognition* (3rd ed.). Academic Press. https://www.sciencedirect.com/book/9780123695314/pattern-recognition

Tobler, C. (2008). *Limits and potential of the concept of indirect discrimination*. Office for Official Publications of the European Communities. https://doi.org/10.2767/56607

Utrecht University. (2024). *Utrecht University developed performance review: "Structural evaluation of AI is needed" - News - Utrecht University*. Utrecht Universtity / News. https://www.uu.nl/en/news/utrecht-university-developed-performance-review-structural-evaluation-of-ai-is-needed

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. https://doi.org/10.1177/13882627211031257

van Bekkum, M., & Zuiderveen Borgesius, F. (2023). Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception? *Computer Law & Security Review*, *48*, 105770. https://doi.org/10.1016/J.CLSR.2022.105770

van Bruxvoort, X., & van Keulen, M. (2021). Framework for Assessing Ethical Aspects of Algorithms and Their Encompassing Socio-Technical System. *Applied Sciences 2021, Vol. 11, Page 11187*, *11*(23), 11187. https://doi.org/10.3390/APP112311187

van der Gaag, S., & Ersoy, S. (2023, June 23). *DUO mag algoritme niet gebruiken totdat meer bekend is over mogelijke discriminatie*. NOS Op 3. https://nos.nl/op3/artikel/2480024-duo-mag-algoritme-niet-gebruiken-totdat-meer-bekend-is-over-mogelijke-discriminatie

van der Sloot, B., Keymolen, E., Noorman, M., Weerts, H. J. P., Wagensveld, Y., & Visser, B. (2021). *Handreiking non-discriminatie by design* . https://research.tue.nl/nl/publications/handreiking-non-discriminatie-by-design

Veldman, A. (2021). Gelijke behandeling. In F. J. L. Pennings & S. S. M. Peters (Eds.), *Europees Arbeidsrecht* (Issue 5, pp. 151–196). Wolters Kluwer. https://dspace.library.uu.nl/handle/1874/415021

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings - International Conference on Software Engineering*, 1–7. https://doi.org/10.1145/3194770.3194776

von Eschenbach, W. J. (2021). Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy and Technology*, *34*(4), 1607–1622. https://doi.org/10.1007/S13347-021-00477-0

Wachter, S., Mittelstadt, B., & Russell, C. (2020). Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.3547922

Wachter, S., Mittelstadt, B., & Russell, C. (2021). Bias Preservation in Machine Learning: The Legality of Fairness Metrics under EU Non-Discrimination Law. *West Virginia Law Review*, *123*(3), 735–790. https://heinonline.org/HOL/Page?handle=hein.journals/wvb123&id=765&div=26&collection=journals

Wang, A., Ramaswamy, V. V., & Russakovsky, O. (2022). Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation. *ACM International Conference Proceeding Series*, *14*, 336–349. https://doi.org/10.1145/3531146.3533101

Wang, R., Harper, F. M., & Zhu, H. (2020, April 21). Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3313831.3376813

Wang, X., Zhang, Y., & Zhu, R. (2022). A brief review on algorithmic fairness. *Management System Engineering 2022 1:1*, *1*(1), 1–13. https://doi.org/10.1007/S44176-022-00006-Z

Watkins, E. A., McKenna, M., & Chen, J. (2022). *The four-fifths rule is not disparate impact: a woeful tale of epistemic trespassing in algorithmic fairness*. https://arxiv.org/abs/2202.09519v1

Weerts, H., Xenidis, R., Tarissan, F., Olsen, H. P., & Pechenizkiy, M. (2023). Algorithmic Unfairness through the Lens of EU Non-Discrimination Law: Or Why the Law is not a Decision Tree. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 805–816. https://doi.org/10.1145/3593013.3594044

Whitworth, B. (2008). A Brief Introduction to Sociotechnical Systems. In M. Khosrow-Pour (Ed.), *Encyclopedia of Information Science and Technology* (2nd ed., Vol. 1, pp. 394–400). IGI Global. https://doi.org/10.4018/978-1-60566-026-4.CH066

Xendis, R., & Senden, L. (2020). EU non-discrimination law in the era of artificial intelligence : mapping the challenges of algorithmic discrimination. In U. Bernitz, X. Groussot, J. Paju, & S. de Vries (Eds.), *General Principles of EU Law and the EU Digital Order* (pp. 267–296). Kluwer Law International. https://cadmus.eui.eu/handle/1814/65845

Xenidis, R. (2021). Tuning EU equality law to algorithmic discrimination: Three pathways to resilience. *Https://Doi.Org/10.1177/1023263X20982173*, *27*(6), 736–758. https://doi.org/10.1177/1023263X20982173

Yu, R., Lee, H., & Kizilcec, R. F. (2021). Should College Dropout Prediction Models Include Protected Attributes? *L@S 2021 - Proceedings of the 8th ACM Conference on Learning @ Scale*, 91–100. https://doi.org/10.1145/3430895.3460139

Zhang, L., Wu, Y., & Wu, X. (2017). *A Causal Framework for Discovering and Removing Direct and Indirect Discrimination*. https://doi.org/10.48550/arXiv.1611.07509

Zhu, Z., Yao, Y., Sun, J., Li, H., & Liu, Y. (2023). Weak Proxies are Sufficient and Preferable for Fairness with Missing Sensitive Attributes. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 43258–43288). PMLR. https://proceedings.mlr.press/v202/zhu23n.html

Zinda, N. (2022). *Ethics Auditing Framework for Trustworthy AI: Lessons from the IT Audit Literature*. 183–207. https://doi.org/10.1007/978-3-031-09846-8_12

Legal sources (laws, directives, regulations, court judgements, etc.)

AI Act: *full name*: Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts - Analysis of the final compromise text with a view to agreement; *year:* 2024; *document type:* compromise (no formal legal status); *document number:* ST 5662 2024 INIT - NOTE; *URL:* https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf

AWGB: *full name:* Algemene wet gelijke behandeling; *year:* 2019 (current version) / 1993 (first version); *document type:* Dutch law; *Identification code:* BWBR0006502; *URL:* https://wetten.overheid.nl/jci1.3:c:BWBR0006502&z=2020-01-01&g=2020-01-01

De Weerd: *full name:* Judgment of the Court (Sixth Chamber) of 24 February 1994. - M. A. De Weerd, née Roks, and others v Bestuur van de Bedrijfsvereniging voor de Gezondheid, Geestelijke en Maatschappelijke Belangen and others. - Reference for a preliminary ruling: Raad van Beroep 's-Hertogenbosch - Netherlands. - Equal treatment for men and women - Social security - Directive 79/7/EEC - Effects of late transposition on rights acquired under the Directive. - Case C-343/92.; *year:* 1994; *document type:* Judgement *document number:* 61992CJ0343; *URL:* https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:61992CJ0343

Directive 2000/43/EC: *full name*: Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin; *year:* 2003 (current version) / 2000 (first version); *document type:* EU Council directive; *document number:* 32000L0043; *URL:* http://data.europa.eu/eli/dir/2000/43/oj

Directive 2000/78/EC: *full name*: Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation; *year:* 2003 (current version) / 2000 (first version); *document type:* EU Council directive; *document number:* 32000L0078; *URL:* http://data.europa.eu/eli/dir/2000/78/oj

Directive 2010/41/EC: *full name*: Directive 2010/41/EU of the European Parliament and of the Council of 7 July 2010 on the application of the principle of equal treatment between men and women engaged in an activity in a self-employed capacity and repealing Council Directive 86/613/EEC; *year:* 2012 (current version) / 2010 (first version); *document type:* EU directive; *document number:* 32010L0041; *URL:* http://data.europa.eu/eli/dir/2010/41/oj

Directive 2004/113/EC: *full name:* Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services); *year:* 2007 (current version) / 2004 (first version); *document type:* EU Council directive; *document number:* 32004L0113; *URL:* http://data.europa.eu/eli/dir/2004/113/oj

Directive 2006/54/EC: *full name*: Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast); *year:* 2008 (current version) / 2006 (first version); *document type:* EU directive; *document number:* 32006L0054; *URL:* http://data.europa.eu/eli/dir/2006/54/oj

GDPR: *full name*: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation); *year:* 2018 (current version) / 2016 (first version); *document type:* EU Council Regulation; *document number:* 32016R0679; *URL:* http://data.europa.eu/eli/reg/2016/679/oj

Gerster: *full name:* Judgment of the Court (Sixth Chamber) of 2 October 1997. - Hellen Gerster v Freistaat Bayern. - Reference for a preliminary ruling: Bayerisches Verwaltungsgericht Ansbach - Germany. - Equal treatment for men and women - Public servant - Part-time employment - Calculation of length of service. - Case C-

1/95.; *year:* 1997; *document type:* judgement; *document number:* 61995J0001; *URL:* https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:61995CJ0001

GW: *full name:* Grondwet voor het Koninkrijk der Nederlanden; *year:* 2023 (current version)/ 1815 (first version); *document type:* Dutch law; *Identification code:* BWBR0001840; *URL:* https://wetten.overheid.nl/BWBR0001840/2023-02-22/0; https://open.overheid.nl/documenten/ronl-bfa3c90611c54fd791d929d4b94e0869ee6a7065/pdf

Parris: *full name:* Reference for a preliminary ruling from The Labour Court, Ireland (Ireland) made on 13 August 2015 — Dr David L. Parris v Trinity College Dublin, Higher Education Authority, Department of Public Expenditure and Reform, Department of Education and Skills.; *year:* 2015; *document type:* Reference for a preliminary ruling; *document number:* 62015CN0443; *URL:* https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62015CN0443

RV: *full name:* Wetboek van Burgerlijke Rechtsvordering; *year*: 1837 (first version) / 2021 (current version); *document type:* Dutch law; *Identification code:* BWBR0001827; *URL:* https://wetten.overheid.nl/jci1.3:c:BWBR0001827&z=2024-01-01&g=2024-01-01

Rinner-Kühn: *full name:* Judgment of the Court (Sixth Chamber) of 13 July 1989. Ingrid Rinner-Kühn v FWW Spezial-Gebäudereinigung GmbH & Co. KG. Reference for a preliminary ruling: Arbeitsgericht Oldenburg - Germany. Continued payment of wates in the event of illness - Exclusion of part-time workers - Article 119 of the EEC Treaty. Case 171/88.; *year:* 1989; *document type:* Judgement *document number:* 61988CJ0171; *URL:* https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:61988CJ0171

Seymour-Smith: *full name:* Judgment of the Court of 9 February 1999. - Regina v Secretary of State for Employment, ex parte Nicole Seymour-Smith and Laura Perez. - Reference for a preliminary ruling: House of Lords - United Kingdom. - Men and women - Equal pay - Equal treatment - Compensation for unfair dismissal - Definition of 'pay' - Right of a worker not to be unfairly dismissed - Whether falling under Article 119 of the EC Treaty or Directive 76/207/EEC - Legal test for determining whether a national measure constitutes indirect discrimination for the purposes of Article 119 of the EC Treaty - Objective justification. - Case C-167/97; *year:* 1999; *case number:* C-167/97; document *type:* judgement; *document number:* 61997CJ0167; *URL:* https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:61997CJ0167

Staat der Nederlanden: *year:* 2020; *document type:* Judgement; *Ecli number:* ECLI:NL:RBDHA:2020:865; *URL:* https://deeplink.rechtspraak.nl/uitspraak?id=ECLI:NL:RBDHA:2020:865

Verdict Belastingdienst/Toeslagen: *year:* 2023; *verdict number: 2023-103*; *claimant:* anonymous; *defendant*: Belastingdienst/Toeslagen *file number*: 2020-0628; URL*:* https://oordelen.mensenrechten.nl/oordeel/2023-103

Villar Láiz; *full name:* Judgment of the Court (Third Chamber) of 8 May 2019. Violeta Villar Láiz v Instituto Nacional de la Seguridad Social (INSS) and Tesorería General de la Seguridad Social (TGSS). Request for a preliminary ruling from the Tribunal Superior de Justicia de Castilla y León. Reference for a preliminary ruling — Equal treatment for men and women in matters of social security — Directive 79/7/EEC — Article 4 — Prohibition of any discrimination on the ground of sex — Indirect discrimination — Part-time work — Calculation of retirement pension. Case C-161/18.; *year:* 2019; *document type:* Judgement *document number:* 62018CJ0161; *URL:* https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62018CJ0161

Voß; *full name:* Judgment of the Court (First Chamber) of 6 December 2007. Ursula Voß v Land Berlin. Reference for a preliminary ruling: Bundesverwaltungsgericht - Germany. Article 141 EC - Principle of equal pay for men and women - Civil servants - Overtime - Indirect discrimination against women employed part-time. Case C-300/06.; *year:* 2007; *document type:* Judgement *document number:* 62006CJ0300; *URL:* https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:62006CJ0300

WCRM: *full name:* Wet College voor de rechten van de mens; *year*: 2019 (current version)/ 2011 (first version); *document type:* Dutch law; *Identification code:* BWBR0030733; *URL:* https://wetten.overheid.nl/jci1.3:c:BWBR0030733&z=2016-01-18&g=2016-01-18

WGBH/CZ: *full name:* Wet gelijke behandeling op grond van handicap of chronische ziekte; *year*: 2019 (current version)/ 2003 (first version); document *type:* Dutch law; *Identification code:* BWBR0014915; *URL:* https://wetten.overheid.nl/jci1.3:c:BWBR0014915&z=2020-01-01&g=2020-01-01

WGBLA: *full name:* Wet gelijke behandeling op grond van leeftijd bij de arbeid; *year*: 2023 (current version)/ 2003 (first version); document *type:* Dutch law; *Identification code:* BWBR0016185; *URL:* https://wetten.overheid.nl/jci1.3:c:BWBR0016185&z=2023-07-01&g=2023-07-01

WGBMV: *full name:* Wet gelijke behandeling van mannen en vrouwen; *year*: 2014 (current version)/ 1980 (first version); document *type:* Dutch law; *Identification code:* BWBR0003299; *URL:* https://wetten.overheid.nl/jci1.3:c:BWBR0003299&z=2015-07-01&g=2015-07-01

WOA: *full name:* Wijzigingswet Burgerlijk Wetboek en Ambtenarenwet ivm verbod tot maken van onderscheid tussen werknemers naar arbeidsduur; *year*: 2011 (current version)/ 1980 (first version); *document type:* Dutch law; *Identification code:* BWBR0008161; *URL:* https://wetten.overheid.nl/jci1.3:c:BWBR0008161&z=2012-10-01&g=2012-10-01

WOBOT: *full name:* Uitvoeringswet EU-richtlijn 1999/70/EG (raamovereenkomst door het EVV, de UNICE en het CEEP inzake arbeidsovereenkomsten voor bepaalde tijd); *year*: 2011 (current version)/ 2002 (first version); *document type:* Dutch law; *Identification code:* BWBR0014195; *URL:* https://wetten.overheid.nl/jci1.3:c:BWBR0014195&z=2012-10-01&g=2012-10-01

YS v. NK: *full name:* Opinion of Advocate General Kokott delivered on 7 May 2020. YS v NK. Request for a preliminary ruling from the Landesgericht Wiener Neustadt. Reference for a preliminary ruling – Equal treatment in employment and occupation – Directives 2000/78/EC and 2006/54/EC – Scope – Prohibition of indirect discrimination on grounds of age or sex – Justifications – National legislation providing for an amount to be withheld from pensions paid directly to their recipients by undertakings in which the State has a majority participation and for the cancellation of the indexation of the amount of those pensions – Articles 16, 17, 20 and 21 of the Charter of Fundamental Rights of the European Union – Applicability – Discrimination on grounds of property – Infringement of the freedom of contract – Infringement of the right to property – Article 47 of the Charter of Fundamental Rights – Right to an effective remedy. Case C-223/19.; *year:* 2020; *document type:* Opinion of the Advocate General; *document number:* 62019CC0223; *URL:* https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:62019CC0223

Appendices

Appendix A: mathematical details

The relationship between acceptance rates and non-acceptance rates

Let G be any group of persons (e.g. a protected group, a comparator group or even the whole population) and let $R_G^+ = \frac{\#favoured\_members\_of\_G}{\#all\_members\_of\_G}$ and $R_G^- = \frac{\#unfavoured\_members\_of\_G}{\#all\_members\_of\_G}$ respectively be the selection rate and non-selection rate of G. Given that every person in G is either favoured (a member of the favourable group) or disadvantaged (a member of the unfavourable group) and no one can be neither or both, it follows that $\#all\_members\_of\_G = \#favoured\_members\_of\_G + \#disadvantaged\_members\_of\_G$. From this the following relation between R_{G}^- and R_{G}^+ can be deduced:

$$R_G^- = \frac{\#unfavoured\_members\_of\_G}{\#all\_members\_of\_G} = \frac{\#all\_members\_of\_G - \#favoured\_members\_of\_G}{\#all\_members\_of\_G}$$

$$= 1 - \frac{\#favoured\_members\_of\_G}{\#all\_members\_of\_G} = 1 - R_G^+.$$

A prove of the equivalence of comparing acceptance rates and comparing non-acceptance rates for comparisons in terms of difference.

Let R_P^+ and R_C^+ respectively be the acceptance rates of a protected group and comparator group and let R_P^- and R_C^- respectively be the non-acceptance rates of this protected group and comparator group. (In both *Seymour-Smith* and *Voß* women are the protected group and men are the comparator group). Evidently, showing that R_P^+ is considerably smaller than R_C^+ (as attempted in *Seymour-Smith*) is the same as showing that R_C^+ is considerably larger than R_P^+, which, if we assume (non-)selection rates should be compared by considering their difference, is the same as showing that $R_C^+ - R_P^+$, is considerably high. Showing that R_P^- is considerably higher than R_C^- (as proposed in *Voß*) on the other hand, is, in terms of difference, the same as showing that $R_P^- - R_C^-$, is considerably high. However, if we assume that everyone in the comparator group and protected group is either advantaged or disadvantaged (and not neither or both), we have shown that that $R_P^- = 1 - R_P^+$ and $R_C^- = 1 - R_C^+$. Hrence, we can deduce that $R_P^- - R_C^- = 1 - R_P^+ - (1 - R_C^+) = R_C^+ - R_P^+$, from which we can conclude that showing that R_P^+ is considerably smaller than R_C^+, in terms of difference, is equivalent to showing that R_P^- is considerably higher than R_C^-, in terms of difference.

Appendix B: Interview excerpts

A.1: "The three protected attributes we used… those are also the ones that were most available. Because there were also some others which are, legally speaking, just as important, but which are a bit harder to gather, like sexual orientation, religious beliefs, et cetera. So, we started with what we had [available]."

B.1: "How do you determine these groups for example? We did that too often in consultation with the business [stakeholders] (…) and also based on what we saw in the data for example. (…) In our methodology we actually started with drawing up a very broad list of all things on the basis of which we did not want to discriminate in principle. And next we had to look: where do we have data on? I think there were five things we, within the municipality, did have data about (…) and for a lot of things we did not."

D.1: "So what we often do is trying to create a summary of prominent publications or, like, trying to find best practices somewhere. So, then we are talking about The Netherlands Institute for Human Rights. Then we are talking about the Court of Audit [of the Netherlands]. Those are the kind of sources we use. But also, simply the law. So, whatever we can find. For example, what is, in fact, protected personal data according to law and things like that? So, we map that and then we also show: okay so to perform a bias test you should have groups [to compare], but what data do you have?"

C.1: "It was quite clear already who, in general, were disadvantaged. (…) The groups that were already largely disadvantaged according to the reports that were already there. But in the commission [executing the audit] itself for each [input] variable it was investigated who exactly could be disadvantaged [by using it]."

D.2: "Sometimes you start with the sensitive feature: okay what group could possibly result from this? And if there are multiple features: what groups, together, combined and sometimes you also just ask a domain expert -or well, we want to include them in this step anyway- like: okay what are the possible risks you see for this type of algorithm and that also often results in groups which are intersectional. (…) So that is very complicated because, yeah, ideally you would compare those intersectional groups, but what's possible with the data is just very limited."

B.2: "I think it could be an interesting research project to look at how you might be able to approach the things [protected attributes] we might not directly have available anyway, using proxy variables. Although one could also argue that those things [protected attributes] are not registered for a certain reason. [Think of] sexual orientation or political opinion, you name it. Then you can question whether you should want it: to approach this [sensitive information] anyway."

B.3: "What we found especially hard is that… if you find a bias on one of those features of which you think that it might relate to a sensitive attribute (…) so proxy variables, if you find a bias on those… to what extent can you assume then that there really is a bias on those underlying attributes or is it just -so to speak- [caused by] the differentiating power of that feature you use. (…) You might be able to delve into that statistically, like: how strong then is that relation between your feature and that attribute, but yeah, at least in our case it still felt like a lot of guesswork (…) Imagine… well… that I find [out] that a distinction is made based on the number of children [a person has] or something like that, does that translate to an actual difference between people with (…) different nationalities for example, because different nationalities have more or less children on average."

D.3: "For example, if we say: 'well, we do not want that our model discriminates different nationalities' we look at to which extent do you have data about nationality or maybe we should consider another possibility, like a proxy. But a proxy is also risky. So that is also the question we ask people who work there [at the organisations we consult in assessing algorithmic fairness] like: are there even proxies of which you can say that they really have a good link or a very strong relation with nationality, for example?"

A.2: (Asked whether they included intersectional groups in their analysis:) "No we did not do that. And that is indeed a very interesting [option] and it is on our radar."

B.4: (Asked whether they included intersectional groups in their analysis:) "No, we actually did not do that for two reasons. Firstly, because it was not indicated by the business [experts] so to speak. So, we sat with them like, well, which groups do you think are important and [of which groups], so to speak, would you suspect from your own experience of the working practice that there could be a difference [in treatment]. Nothing like that [(intersectionality)] really came up there actually. The second reason is the lack of data. Of course, you need a certain group size to be able to say something meaningful and as soon as you start combining groups or start combining features that [(having a sufficient group size)] often is not the case anymore. "

C.2: "You can look at fairness in multiple ways actually, because fairness has multiple notions. There are several ways of approaching fairness – or actually rather equality, if you define it like that. And if you look at how you want to prevent non-discrimination [sic] then, then you can look at how you treat different groups, for example. You can treat them equally, but you can also treat them a certain way within those groups: (…) [meaning] that you do not only have that people's scores are literally evened, but that you look at how you can favour certain groups so that they will be equal to the other or something like that. (…) And so, there are many forms of this. (…) I just hear way too little of this conversation, like: which notion [of fairness] should you actually use in what context? There is a complete ethical deliberation behind that, like: you cannot just pick a notion. You need to think about how then you knowingly want to favour people to create that fairness, so that is really a very hard conversation."

B.5: "When it comes to [fairness] metrics we have benefited a lot from (…) the fairness decision tree. (…) Then we ended up somewhere down [the fairness decision tree] with three metrics left, I believe and then we really tried to understand, to look, what [metric] was most fitting. And at a certain point we submitted that to our stakeholders from the business. So, in fact we approached it in such a way that we did a proposal and asked them [the stakeholders] feedback, or approval of it."

D.4: "We always think the fairness tree is a nice source. We adapted it once. [In its original form] you actually need to choose from the very first step (…): do we want that the model is fair based on representation or equal distribution of errors. And then we adapted it so that you actually say: we test for both, always. (…) [we] also included it in a workshop [for a client that wanted to test the fairness of their algorithm] once. And then actually the analysists -each analysist was responsible for an algorithm and there were three of them- they just had to work through this fairness tree as a homework assignment and then let us know the next week what metric they would pick. And then we ended up with that false positves/groupsize parity, I think, so that is like one fairness metric."

D.5: "So analysts (…) are glad when they see a fairness tree, because then they finally see something technical. All those steps before were less technical. But yeah, they end up with a number and they are like: done? Am I done now? (…) And then [we ask] okay, but what does it mean? [And then the analysts say] yeah, I don't know."

C.3: "Because many tools… I do not really know what they assume. (…) I do not know that very well what to think of it, because I do not see the fairness method behind it that well. (…) The one who executes this [fairness assessment] should really be hyper-aware of how this [tool] works, but also that fairness notion."

D.6: "And ideally, we want that this fairness metric would also be chosen by a diverse group (…) But then you really need to make an enormous translation effort into what each fairness metric means. And people just do not always feel like that or have the time."

A.3: "Part of this process [of choosing a suitable fairness metric] was doing literature research together with other data scientists and have a look at: okay, what are different ways of measuring bias? (…) It started like a kind of data science project with a literature review. Then we went completely all out on all kinds of papers with all kinds of complex metrics. (…) Then we implemented a few and then we discovered quite quickly that if we wanted to have a metric that was meaningful within the wider organisation, we just needed something that was simple to explain. So, in the end we chose the very simplest metric. That is a variant of demographic parity.[58] (…) And importantly, I heard that more often later on, that people say: in the end, it should be understandable, especially if you want this to land within the broader organisation indeed."

D.7: "Actually we concluded like: okay, there should be much more guidance in what then such a number [resulting from using a fairness metric] means, what you could do with it, so then we (…) firstly tried to express the number in something else. So, then we said: okay, this number means that, for example, two out of ten men are selected and that (…) six out of ten women are selected, so such a number can be translated to something that is more understandable and then we decided a bit like: what do you think of that? (…) So that, ultimately, whether that is undesirable, that difference [in selection ratio], that that is actually a decision that, for example, management should make (…) But then we help the data analysts make that translation."

B.6: "It is very tempting to say: well, we try to find a hard threshold [for the chosen fairness metric]. And we did try hard to do that, but ultimately, we decided not to do that. (…) we had a working group, or an advice group, I should actually say, also with external people (…) ethical AI researchers for example, all of whom also said like: try not to do that [(using a hard threshold)], because ultimately it is always a subjective decision. And something like the four-fifths rule as well… yeah you find it on the internet, so then it feels objective, but in fact it is still, yeah, just someone who invented that. (…) I think it is principally a political choice where you place the threshold and what is acceptable for you. So, in our case that means, well, that in fact the alderman[59] should decide that for an algorithm"

D.8: "The very idea of a bias analysis is that you actually document the choices of all steps [in the process of algorithm development]. And then you end up with a certain result -for example: one out of ten women is selected against six out of ten men- and [the idea is] that you deliver that whole package to management or, like, a person who has to form an opinion on that… on what kind of follow up actions should be executed. So, for that reason you should document and also document it in such a way that it is also understandable for someone with a less technical background."

A.4: "What we as data scientists did is [we] said like: okay, we can map all these metrics. We can come up with statistics for that. We made a few printouts (…) [and came] with a call to action to a group of business stakeholders with the message like: okay, we can observe this. What should we do with it? Also, from the point of view that we said like: (…) this is not a technical problem anymore. Or this is (…) not a matter of: how are you going to solve this as data scientists, but this is a conversation that we

[58] Demographic parity is also known as statistical parity.

[59] In Dutch the word "wethouder" was used.

should have within the organisation. (…) But in all honesty, I have to I have to admit it ended there. (…) Everyone was super fascinated and found it interesting, but from there it slowly died a silent dead.”

A.5: “At a certain point you as a company should have an opinion about this and that should be, I am afraid, on a CE [(chief executive)] level, so to speak. There someone should say like: okay, this is where we are currently at, but this is where we want to go. And that part never happened. And I think: as soon as (…) someone says, even if it is super high-level, like: this is our ambition; this is what we strive for, then we can implement it and make it smaller again.”

A.6: “As data scientists we are able to say: we have a baseline measurement now. If we are going to develop new [input] features now, we can do another measurement to look at what the impact of new features on bias metrics is.”

B.7: “But how we worked around it [(the problems of interpreting the value of a fairness metric)] a little is by looking at: how are things in comparison with the current [non-algorithmic] process. (…) It is hard to say objectively: ‘we think this is good enough or not’, but it is easier to say: ‘at least it is better than what happens now’.”

A.7: (Asked where their training data originated from and whether they knew whether the training data was representative for the target population:) “We do hold that assumption, that it [(the training data)] is representative. So, the training data for our system are placements that were made within [name of organisation] (…)”

B.8: (Asked whether they looked at the representativeness of the training and evaluation data:) “No, we did not look at that. I would expect it is [representative], because it is trained on data from the working process, in principle, without filters, so to speak, only the filters (…) [selecting the people] where the model will be used for.”

A.8: “(…) it is important to note that those placements[60]… they result from our business processes, not necessarily from the recommender system itself. (…) We have many small offices in the country. It could be that a job seeker enters an office and starts talking to an intermediary and the intermediary says: ‘hey, wait, I know a good job for you. I will introduce you to that client.’ And than ultimately someone gets placed. That is all input for our algorithm. And with that we hope to ensure that the algorithm, so to speak, gets to see training data or gets to see labels outside of his or her own reality.”

A.9: “If there is a placement coming completely from outside of the [algorithmic recommendation] system, then it could well be that the information that we have about the jobseeker is not sufficient [to include it in the training data].”

A.10: “We did not do that (…) At that time we also scoped it a bit like: ‘We are data scientists. We are going to focus on the algorithm. We are going to show that and then we are going to ask the organisation like: what should we further look at?’”

D.9: “We looked indeed at the processes they already executed and there we also saw that they already did representativeness tests, so as part of their data quality activities, they also looked at representativeness already. And then we did explain that further in a few workshops, like the importance of (…) that training data, that it is representative (…) for the purpose that you ultimately want to achieve with the algorithm. (…)”

D.10: “And we also looked at: well, okay, if you perform a bias analysis, then you do that on a test set for example. Is that test set on its turn representative of the whole training data set and if not: what is

[60] See excerpt A.7 for the meaning of placements in this context.

needed [for that]? Is it the size of the test set, for example? Is it the distribution of the [protected] groups within the test set?”

A.11: (Asked whether there were plans to periodically retest or monitor fairness:) “No, so those existed, but they do not exists anymore, because (…) we are in a transition within our organisation and as a part of that transition eventually the plug is going to be pulled on the algorithm. We already know that. But it is the ambition to continue this [(assessing fairness)] within the new systems we are developing. That’s for sure.”

B.9: “So there is a management plan… Assuming that the pilot will turn out to be successful and it will really be implemented, [the idea is] to do it [(the bias analysis)] yearly. And there will be (…) a person appointed for that who coordinates that [and] who subsequently has to form a working group every year, not only to perform the bias analysis another time, but to check a whole lot of things actually, like: can the features still pass muster, (…) have there not been adjustments to the law for example, because of which the model can or cannot take certain things [into consideration] anymore, et cetera.”

D.11: “We developed the whole bias analysis together with them [(the organisation which we consulted)] and we looked at what they did in their monitoring phase. So, in that algorithm life cycle they have a monitoring phase and there we (…) added a few tests. So I think we added a few questions that were qualitative of nature (…) and a few quantitative tests, of which we said like: okay, if you applied a certain fairness metric during the bias analysis, then you should monitor it periodically as well, so then you should look at for example whether you want to perform that (…) once in every two months or yearly, depends [on] what kind of algorithm you have. (…) We really always attach that bias analysis to the monitoring phase as well so that you always keep looking there indeed at the data, for example, at whether the model still does what it should do. That is what often is not seen enough. It is not a one-off exercise.”

Appendix C: The Fairness Tree from the Aequitas Toolkit

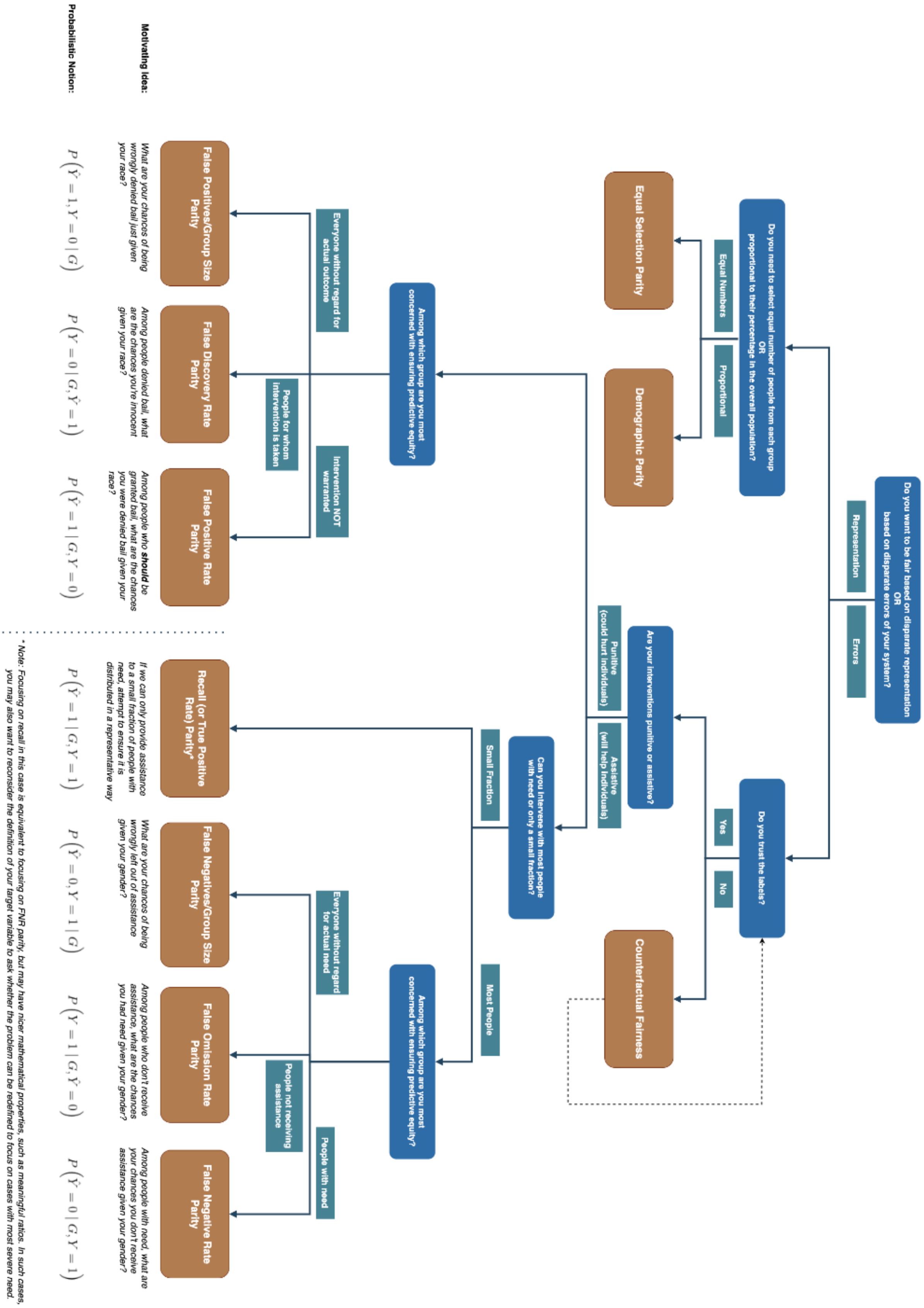

Do you want to be fair based on disparate representation OR based on disparate errors of your system?
Representation
Errors
Do you need to select equal number of people from each group OR proportional to their percentage in the overall population?
Equal Numbers
Proportional
Equal Selection Parity
Demographic Parity
Do you trust the labels?
Yes
No
Counterfactual Fairness
Are your interventions punitive or assistive?
Punitive (could hurt individuals)
Assistive (will help individuals)
Among which group are you most concerned with ensuring predictive equity?
Everyone without regard for actual outcome
People for whom intervention is taken
Intervention NOT warranted
Can you intervene with most people with need or only a small fraction?
Small Fraction
Most People
Among which group are you most concerned with ensuring predictive equity?
Everyone without regard for actual need
People not receiving assistance
People with need
False Positives/Group Size Parity
False Discovery Rate Parity
False Positive Rate Parity
Recall (or True Positive Rate) Parity*
False Negatives/Group Size Parity
False Omission Rate Parity
False Negative Rate Parity
Motivating Idea:
What are your chances of being wrongly denied bail just given your race?
Among people denied bail, what are the chances you're innocent given your race?
Among people who should be granted bail, what are the chances you were denied bail given your race?
If we can only provide assistance to a small fraction of people with need, attempt to ensure it is distributed in a representative way
What are your chances of being wrongly left out of assistance given your gender?
Among people who don't receive assistance, what are the chances you had need given your gender?
Among people with need, what are your chances you don't receive assistance given your gender?
Probabilistic Notion:
$P(\hat{Y}=1, Y=0 \mid G)$
$P(Y=0 \mid G, \hat{Y}=1)$
$P(\hat{Y}=1 \mid G, Y=0)$
$P(\hat{Y}=1 \mid G, Y=1)$
$P(\hat{Y}=0, Y=1 \mid G)$
$P(Y=1 \mid G, \hat{Y}=0)$
$P(\hat{Y}=0 \mid G, Y=1)$
* Note: Focusing on recall in this case is equivalent to focusing on FNR parity, but may have nicer mathematical properties, such as meaningful ratios. In such cases, you may also want to reconsider the definition of your target variable to ask whether the problem can be redefined to focus on cases with most severe need.

Figure 4: The Aequitas Fairness Tree. *This tree is meant as a tool to select the fairness metric that is most appropriate for the use context of an algorithm. See Saleiro et al. (2019) for more information on its intended use.*