Voorwoord Hans Frankenprijs 2023

In 1989 is de Stichting Geschillenoplossing Automatisering (SGOA)[1] opgericht om te helpen bij het oplossen en voorkomen van ICT-conflicten. De SGOA treedt hierbij op als onafhankelijke instantie waarbij partijen kunnen aankloppen voor deskundige arbiters, mediators en deskundigen. In de afgelopen jaren heeft de SGOA kunnen helpen met het oplossen van allerlei ICT-conflicten: stroef lopende projecten, geschillen over auteursrecht en eigendom maar ook vragen over digitale data, informatiebeveiliging, privacy en algoritmes.

Naast het uitvoeren van haar kerntaak als geschilleninstantie vindt de SGOA het belangrijk de maatschappelijke aandacht voor het raakvlak van recht en digitale technologie te stimuleren. Daarom heeft het bestuur van de SGOA in 2014 de Hans Frankenprijs in het leven geroepen. Deze prijs is vernoemd naar emeritus-hoogleraar Hans Franken, die gedurende vele jaren als bestuursvoorzitter het boegbeeld van de SGOA was. Met deze prijs wil de SGOA studenten en hun begeleiders stimuleren tot het schrijven van een scriptie op het gebied van IT en recht, en daarmee de ontwikkeling van kennis over juridische aspecten van ICT bevorderen. De prijs wordt elke twee jaar uitgereikt, in 2023 voor de vijfde maal.

Een multidisciplinaire jury, samengesteld uit de kring van de aan de SGOA verbonden deskundigen, heeft daartoe met groot plezier een groot aantal lezenswaardige scripties op het gebied van IT en recht beoordeeld. De scripties waren afkomstig van diverse Nederlandse universiteiten en hogescholen.

[1] https://sgoa.eu/

Dit jaar had de jury een moeilijke taak om uit meerdere uitstekende scripties de beste te moeten kiezen. Na intensief overleg heef de jury besloten de Hans Frankenprijs in 2023 toe te kennen aan de scriptie van Robin Verhoef, getiteld: "Possibilities and Problems of Combating Discrimination in Automated Decision-Making with Technical Methods in an EU law context". Het in deze scriptie beschreven onderzoek is uitgevoerd aan de Tilburg University, onder begeleiding van Charmian Lim.

De scriptie behandelt het probleem van het herkennen en voorkomen van discriminatie in besluiten genomen door algoritmes. Dit is een zeer actueel probleem, vanwege de snelle ontwikkelingen op het gebied van kunstmatige intelligentie. Het bijzondere is dat de scriptie uit zowel een juridisch als een technisch deel bevat. Geheel in lijn met de principes van de SGOA wordt er niet alleen gekeken naar juridische eisen, maar ook naar technische mogelijkheden.

De scriptie opent met een bespreking van relevante wetgeving op het gebied van discriminatie, gebruik persoonsgegevens en toepassing van AI. De conclusie is dat bepaalde vormen van discriminatie verboden zijn en dat organisaties die algoritmes gebruiken dus stappen moeten nemen om te zorgen dat zij niet de wet overtreden. In het technische deel wordt vervolgens gekeken hoe discriminatie kan worden gemeten en ook deels worden weggenomen. Met behulp van een voorbeeld-dataset wordt het principe van 'data-massage' getoond, een manier om ongewenste discriminatie te verminderen. Hiermee wordt ook een praktische bijdrage geleverd die organisaties helpt om ook daadwerkelijk stappen te nemen om algoritmes meer verantwoord toe te passen.

De combinatie van de twee aanpakken resulteert in een zeer relevante, lezenswaardige scriptie die de meerwaarde duidelijk maakt van multidisciplinair onderzoek. Door informatica-kennis en juridische kennis te combineren ontstaan mogelijkheden om nieuwe technologie op een verantwoorde manier toe te passen. Hierdoor kunnen

toekomstige conflicten over algoritmes worden voorkomen.

De jury van de Hans Frankenprijs 2023 is daarom erg blij met de concrete bijdrage die deze scriptie levert aan het debat over de toepassing van algoritmes en automatische besluitvorming, en hoopt dat deze scriptie breed gelezen wordt en bijdraagt aan het meten en voorkomen van discriminatie in de praktijk. De jury feliciteert Robin van harte met het winnen van deze meer dan welverdiende prijs.

Heemstede, mei 2023
De Jury van de Hans Frankenprijs 2023

Foreword Hans Franken Prize 2023

The Foundation for the Settlement of Automation Disputes (SGOA)[2] was founded in 1989 to help resolve and prevent ICT disputes. In doing so, the SGOA acts as an independent body to which parties can turn for expert arbitrators, mediators and experts. In recent years, the SGOA has been able to help resolve all kinds of ICT conflicts: stalled projects, copyright and ownership disputes but also questions about digital data, information security, privacy and algorithms.

In addition to carrying out its core task as a dispute resolution body, the SGOA considers it important to promote public awareness of the interface of law and digital technology. This is why the SGOA board created the Hans Franken Prize in 2014. This prize is named after emeritus professor Hans Franken, who was the figurehead of the SGOA for many years as chairman of the board. With this prize, the SGOA wants to encourage students and their supervisors to write a thesis in the field of IT and law, thus promoting the development of knowledge on legal aspects of ICT. The prize is awarded every two years, the fifth time in 2023.

To this end, a multidisciplinary jury composed of the circle of experts associated with the SGOA took great pleasure in assessing a large number of readable theses in the field of IT and law. The theses came from various Dutch universities and colleges.

This year, the jury had the difficult task of having to choose the best from several excellent theses. After intensive deliberation, the jury decided to award the Hans Franken Prize in 2023 to Robin Verhoef's thesis entitled: "Possibilities and Problems of Combating Discrimination in Automated Decision-Making with Technical Methods in an EU law context". The research described in this thesis was conducted at Tilburg University, under the supervision of Charmian Lim.

[2] https://sgoa.eu/

The thesis addresses the problem of recognising and preventing discrimination in decisions made by algorithms. This is a very topical problem because of the rapid developments in artificial intelligence. What makes the thesis special is that it consists of both a legal and a technical part. Fully in line with the principles of the SGOA, it not only looks at legal requirements but also at technical possibilities.

The thesis opens with a discussion of relevant legislation on discrimination, use of personal data and application of AI. It concludes that certain forms of discrimination are prohibited and so organisations using algorithms should take steps to ensure they do not break the law. The technical section then looks at how discrimination can be measured and also partly eliminated. Using an example dataset, the principle of 'data-massage' is shown, a way to reduce unwanted discrimination. This also provides a practical contribution that helps organisations actually take steps to apply algorithms more responsibly.

The combination of the two approaches results in a highly relevant, readable thesis that highlights the added value of multidisciplinary research. Combining computer science knowledge and legal knowledge creates opportunities to apply new technology responsibly. This can prevent future conflicts over algorithms.

The jury of the Hans Franken Prize 2023 is therefore very pleased with the concrete contribution this thesis makes to the debate on the application of algorithms and automatic decision-making, and hopes that it will be widely read and contribute to the measurement and prevention of discrimination in practice. The jury warmly congratulates Robin on winning this more than well-deserved prize.

Heemstede, May 2023
The Jury of the Hans Franken Prize 2023

Translation by Robin Verhoef and DeepL.

Possibilities and Problems of Combating Discrimination in Automated Decision-Making with Technical Methods in an EU law context

R.I. Verhoef

Master Thesis

Supervisors: C.D. Lim and A. de Groot

Tilburg Institute for Law, Technology, and Society

Tilburg Law School

2021-2022

Contents

Table of cases

Other Jurisdictions

1 Introduction

1.1 Background

Artificial Intelligence (AI) promises endless possibilities but unfortunately often ends in problems. Possibilities of AI include instant translation between any two languages, the chess computer AlphaZero and digital assistants like Siri and Alexa. These are examples of polished products, built on AI models, which provide users with even quicker access to services. On the dark side of AI, the problems impact real people and affect their lives. When the Dutch Tax and Customs Administration developed a model to detect fraud with certain benefits, the model was found to be discriminating against parents with a dual nationality.[1] Other possible problematic applications of AI are predictions of teacher quality, and predictive policing.[2] When measuring teacher quality, the type of students a teacher teaches and their backgrounds can have a big influence on their success, while totally outside the control of the teacher (and vice versa for students). This means that a teacher will partially be evaluated based on who they are teaching and not how well they teach. In the case of predictive policing, a system which adds more policing to

---

[1] Autoriteit Persoonsgegevens, *De verwerking van de nationaliteit van aanvragers van kinderopvangtoeslag* (2020).

[2] C O'Neil, *Weapons of math destruction: how big data increases inequality and threatens democracy* (Penguin Books 2016).

areas with high crime rates can reinforce its own results since more policing leads to more police action which leads to a higher arrest rate. In both cases, the AI system does not capable of addressing the cause of the issue.

Systems like those I described often discriminate. Discrimination is treating people differently based on an often arbitrary characteristic, so called protected grounds.[3] The exact list of protected grounds varies between different laws, but the list usually includes at least sex, race, skin-colour, religion, political beliefs, and nationality.[4] For example, not hiring someone who does not speak Dutch makes sense in a business with mostly Dutch customers, but not hiring a Belgian who speaks Flemish (a Dutch dialect) for that job is probably illegal discrimination.

The main characteristic of computers (and thus also of AI systems) is repeatability. A computer will not change what it does unless it is instructed to, while people can change their mind and be flexible. This means that the decisions an AI system makes will be repeated, with no room for flexibility. Thus, if an AI system discriminates, it will do so consistently. The problems for AI systems in general are amplified when they appear in an automated individual decision-making system which make decisions which can have a large impact on the people involved. Here, any kind of illegal discrimination in the system will lead to disadvantages for individuals. In systems like these, the prevention of illegal discrimination is paramount for a fair society and to comply with non-discrimination laws.

It might seem like there is an easy solution which prevents illegal discrimination in AI systems: just remove any information based on which you are not allowed to make decisions. A problem with this solution that the discriminatory effects can

---

[3] European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (2018 edition, Publications Office of the European Union 2018) 160.

[4] Convention for the Protection of Human Rights and Fundamental Freedoms (opened for signature 4 November 1950, entered into force 3 September 1953) 213 UNTS 221 (ECHR) 14.

be replicated based on other variables, so called proxy discrimination.[5] Removing attributes is therefore not a fail-safe solution to the problem of illegal discrimination. Another possible solution is to use the protected attributes to check if an AI system discriminates and then correct it. The General Data Protection Regulation (GDPR)[6] does not recognise checking for illegal discrimination as a valid reason for processing data on its own and so this is not yet allowed if the protected attributes are only collected to check for illegal discrimination.[7] The proposed AI Act of the European Commission (AI act) does include a basis for processing protected attributes to check for illegal discrimination, so this might be a future solution.[8] If you would have access to the protected attributes, you can create a system which actively prevents any distinction based on this data in the final AI system.[9]

Next to problems in AI models, it might also be the case that the training data used is problematic. Data is collected in the real world and if it is an accurate representation of the world, any dataset can be filled with discrimination just like the real world. There are historical datasets with clear discriminatory assumptions, like the Boston housing dataset which was collected with the assumption that people want to live in segregated neighbourhoods.[10] While not all datasets are so

---

5 AER Prince and D Schwarcz, 'Proxy Discrimination in the Age of Artificial Intelligence and Big Data' en 105 IOWA LAW REVIEW 62.

6 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) [2016] OJ L119/1 (GDPR).

7 GDPR, art 9.

8 Commission, *Proposal for a Regulation laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)* (COM/2021/206 final) Article 10(5).

9 F Kamiran and T Calders, 'Classifying without discriminating' (2009).

10 M Carlisle, racist data destruction?, 'Medium' (January 2020) <https://medium.com/@docintangible/racist-data-destruction-113e3eff54a8> accessed 7 July 2022.

clearly flawed, they carry the burden with them of the world in which they were collected.

In order to be able to automatically detect discrimination in AI systems, the relevant protected attributes need to be available.[11] If this information is available, then it might be possible to counteract the discrimination in the AI system with certain computational technical methods.[12][13] However, it is not yet clear to which extend these methods are able to remove discrimination and how well they connect to the legal requirements of non-discrimination law and data protection law.[14]

Based on these observations, my main and sub research questions are:

Are there useful technical methods to mitigate illegal discrimination in automated decision-making systems?

(1) What is the legal framework for discrimination within the European Union?

(2) How can illegal discrimination in an automated decision-making system be automatically identified?

(3) What technical methods can combat illegal discrimination?

(4) How useful are the examined technical methods for preventing illegal discrimination?

---

[11] I Žliobaitė and B Custers, 'Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models' (2016) 24(2) Artificial Intelligence and Law 2016 24:2 183.

[12] Žliobaitė and Custers (n 11).

[13] European Union Agency for Fundamental Rights, *'#BigData: Discrimination in data-supported decision making* (2018).

[14] A Balayn and S Gürses, *Beyond Debiasing: Regulating AI and its inequalities* (2021).

1.1.1 Automated decision-making

'Automated decisions are decisions taken using personal data processed solely by automatic means without any human intervention'.[15] Data subjects, which are the people whose data is being processed, should not be subjected to decisions made solely using automated decision-making (ADM), if the decision produces legal effects or similarly significantly affects the data subject.[16] It is not important how the ADM system works, since Article 22 GDPR is concerned with the results of the system. The article contains three circumstances where the use of an ADM system in this context is allowed, so there is no complete prohibition of the use of ADM systems. In all circumstances where ADM systems can be used, the data controller must safeguard the rights and freedoms and legitimate interests of the data subjects.[17] Recital 71 of the GDPR requires the controller of an ADM system to prevent discriminatory effects on natural persons. Thus, non-discrimination law is of special importance of any system to which Article 22 GDPR applies.

1.1.2 Non-discrimination law

The term discrimination has multiple meanings. It means 'the ability to see the difference between two things or people' but also 'treating a person or particular group of people differently, especially in a worse way from the way you treat other people' according to the Cambridge dictionary. While the first describes what we do when making any decision and what AI systems are supposed to do, the second definition can be a problematic consequence. The lawmakers of Europe and the EU

---

[15] European Union Agency for Fundamental Rights, *Handbook on European data protection law* (2018 edition, Publications Office of the European Union 2018) 233.

[16] GDPR, art 22.

[17] GDPR, arts 22(2)(2) and 22(3).

agree that discrimination should be illegal and thus there are a range of European treaties and laws that prohibit discrimination.

Both the European Union (EU) and Council of Europe (CoE) legal orders recognise two types of discrimination, direct and indirect discrimination. Direct discrimination occurs when an individual is treated worse due to a characteristic which falls under a protected ground.[18] Indirect discrimination is more subtle than direct discrimination and occurs when a neutral rule affects a group defined by a protected ground in a significantly more negative way than others in a similar situation.[19] The list with protected grounds is not universal, but it usually includes sex, gender identity, sexual orientation, disability, age, race, ethnic origin, national origin and religion or belief.[20]

1.1.3 Bias

In the computer science literature, wrongful discrimination by a system like an ADM system is called bias.[21] Bias means that a system has a certain systematic level of error. This error might be found in the datasets which are used to develop an ADM system, for example, if there are very few women in the dataset which have a high score.[22] Bias can also be found within the system, for example, if being a woman is counted as a negative attribute, or in the outputs of a system if

---

[18] European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (n 3) 43.

[19] European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (n 3) 54.

[20] European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (n 3) 161.

[21] Balayn and Gürses (n 14).

[22] J Stoyanovich and others, 'Responsible data management' en (2022) 65(6) Communications of the ACM 64.

women are systematically scored lower. The focus of research is on the outputs of systems, since there, bias does the most harm.[23]

There are various ways of measuring the bias in an ADM system. These are called fairness metrics or bias metrics since fairness is considered to be a lack of bias. There are three main levels for measuring bias, on a group level, an individual level, or the causal structure of the system.[24] Each of these approaches focus on different criteria and aim for different goals. This leads to the counter-intuitive situation that the different approaches sometimes do not agree on what a fair outcome is. While there might not be a perfect approach available, choosing one of them will allow the use of technical methods to reduce the bias measured with that specific approach.

1.1.4 Technical methods

Inappropriately designed and trained algorithms can discriminate against certain people and groups.[25] This can be due to the intentions of the maker, but illegal discrimination can also occur when it is not intended, since the data used to train the system might not be correct or a good representation of the world.[26] There are already techniques for discovering illegal discrimination[27] and for removing it once

---

[23] Balayn and Gürses (n 14).

[24] S Verma and J Rubin, 'Fairness definitions explained' (FairWare '18, Association for Computing Machinery May 2018).

[25] Žliobaitė and Custers (n 11).

[26] T Calders and I Žliobaitė, 'Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures' in *Studies in Applied Philosophy, Epistemology and Rational Ethics* (Springer, Berlin, Heidelberg 2013) vol 3.

[27] Žliobaitė and Custers (n 11).

the designer is aware of it.[28] Anti-discrimination techniques can be very useful for ADM systems since illegal discrimination can have a big impact on the people involved. These must do more than just removing variables on which you are not allowed to discriminate, since that is not enough to remove illegal discrimination.[29]

1.2 Gap in the literature

Discrimination is generally undesirable and often illegal. Specifically for ADM systems which fall under Article 22 GDPR, the consequences of discrimination can be very large, since each individual decision made by these ADM systems has a serious impact on a specific data subject. The data controller has an obligation to safeguard the rights and freedoms of data subjects who are subjected to an ADM system in a context where Article 22 GDPR applies and this includes compliance with non-discrimination law. There are technical methods which aim to remove discriminatory effects from systems like ADM systems. Some authors are critical of using a technical approach to combat something like discrimination,[30] and they might be far from perfect. However, currently many companies and organisations are far from GDPR compliant.[31] Technical methods might be a pragmatic way for those who use ADM systems to (partially) comply with their Article 22 GDPR obligations, even if these methods do not function perfectly.

---

[28] F Kamiran, I Žliobaitė, and T Calders, 'Quantifying explainable discrimination and removing illegal discrimination in automated decision making' (2012) 35(3) Knowledge and Information Systems 2012 35:3 613.

[29] Žliobaitė and Custers (n 11).

[30] S Wachter, B Mittelstadt, and C Russell, 'Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI' [2020] (ID 3547922) .

[31] 30% of European businesses are still not compliant with GDPR, 'RSM Global' (July 2019) <https://www.rsm.global/news/30-european-businesses-are-still-not-compliant-gdpr> accessed 16 July 2022.

This research seeks to explore whether technical methods can contribute to anti-discrimination efforts. Specifically, I will analyse the current legal framework and a selection of the available technical methods and test them on an example case study. I will look at how (well) the selected technical methods work in order to determine if they can be useful to mitigate discrimination and comply with non-discrimination law.

1.3 Methodology

To answer my research question, I will combine an analysis of the legal literature about discrimination in ADM systems with an analysis of the technical possibilities which can be found in the computer science literature. The combination of the two will allow me to analyse specific methods and their properties. I will create an example dataset on which to test the methods and evaluate the outcomes. This will allow me to determine if the specific methods are useful to mitigating the illegal discrimination in the ADM system.

1.4 Chapter overview

The second chapter, I will use a literature study to lay out the legal framework for discrimination within the EU. I will look at both overarching, fundamental rights laws and specific EU directives. With this chapter, I will answer sub question one.

In the third chapter, I look at ways to check for and measure discrimination in ADM systems by looking at the literature. After a theoretical description, I create an example dataset with a discriminatory aspect in it which will be used in the following chapters as a case study. The scope of discrimination from Chapter 2 will be used here and extended with the criticisms of the EDRi report on the

standard ways to measure discrimination.[32] With this chapter, I will answer sub question two.

In the fourth chapter, I will use a literature study in combination with technical examinations to look at the state of the art of technical anti-discrimination methods. For this, I will start by looking at the work of Zliobaite, Custers, Calders, and Kamiran.[33] By analysing their methods and applying them to the case study from Chapter 3, I will answer third sub question three.

In the fifth chapter, my aim is to take the results from Chapter 4 and see if these are sufficient to comply with the legal framework laid out in Chapter 2. For this, it is important to look at the technical requirements for each method as well as possible weaknesses and other relevant information like the data a method requires. With this chapter, I will answer sub question four.

In the sixth chapter, the conclusion, I will combine the previous five chapters to synthesize the answer to my main research question. I will also provide a critical reflection on the limitations of the results as well as possible topics for further research.

---

[32] Balayn and Gürses (n 14).

[33] Žliobaitė and Custers (n 11); F Kamiran and T Calders, 'Data preprocessing techniques for classification without discrimination' (2012) 33(1) Knowledge and Information Systems 1

2 Non-discrimination law in the EU

Non-discrimination law aims to protect people from discrimination. Its goal is to ensure that all individuals in a society are treated equally and fairly and have a fair chance at the opportunities of their society.[1] These laws aim to protect people from discrimination in two different situations. Instruments like the European Convention on Human Rights (ECHR)[2] and the Charter of Fundamental Rights (CFR)[3] protect people from states, while the specific anti-discrimination EU directives protect people in specific contexts in their public lives.[4] This body of law provides the minimum level of protection and will be laid out in this chapter. After explaining what is protected, I will continue with laying out the limitations of non-discriminations when applied to ADM and how the GDPR extends the obligations with Article 22 and Recitals 71 and 72.

---

1 European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (2018 edition, Publications Office of the European Union 2018) 42.

2 Convention for the Protection of Human Rights and Fundamental Freedoms (opened for signature 4 November 1950, entered into force 3 September 1953) 213 UNTS 221 (ECHR).

3 Charter of Fundamental Rights of the European Union [2012] OJ C326/391 (CFR).

4 Employment, welfare and social security, education, and access to goods and services. European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (n 1) 109

The EU has multiple laws which prohibit discrimination. The right to equality before the law is considered to be a basic principle of Community Law by the Court of Justice of the European Union (CJEU).[5] The CFR contains the right to 'equality before the law' in Article 20 and the right to non-discrimination in Article 21. These rights apply to any EU law, both on an EU and a national level. Next to these general provisions, the EU also has many directives focused specifically on non-discrimination. The main equality directives are, in alphabetical order, the Employment Equality Directive,[6] the Equal Treatment Directive,
[7] the Gender Goods and Services Directive,[8] and the Racial Equality Directive.[9] These directives comprise the main body of European non-discrimination law and they each require member states to implement protection in the specific area of focus of the directive.

2.1 Direct discrimination

As I already highlighted in Chapter 1, direct discrimination occurs when an individual is treated worse due to a characteristic they hold which falls under a

---

5 European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (n 1) 35.

6 Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation [2000] OJ L303/16.

7 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) [2006] OJ L204/23.

8 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 [2004] OJ L373/37.

9 Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin [2000] OJ L180/22; European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (n 1) 37

protected ground.[10] The list with protected grounds depends on the specific law and it is not universal, but it usually includes sex, gender identity, sexual orientation, disability, age, race, ethnic origin, national origin and religion or belief.[11] Direct discrimination is the most clear-cut type of discrimination and relatively easy to spot. For example, not hiring a woman because she is pregnant is clearly treating her worse than her male counterparts, since the men cannot become pregnant.

To determine if some treatment should be considered direct discrimination, there are three elements to consider.[12] The first element of direct discrimination is that the treatment of a person must be less favourable than what is normal. Usually, this is easy to determine. A person could be denied a service, have their social security benefits revoked or be victim of abuse or violence while others in similar positions are treated better. The European Court of Human Rights (ECtHR) also considers treating two people in widely different situations in the same way a form of less favourable treatment.[13]

The second element is that the less favourable treatment has to be compared to the treatment of another person. This second person, the comparator, should be in similar circumstances as the person who is being discriminated. The main difference between the two should be the possession of a specific protected characteristic. A man and a woman working in similar roles in the same company are suitable comparators for each other.[14] Their circumstances are comparable and they both

---

[10] European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (n 1) 43.

[11] European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (n 1) 49.

[12] European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (n 1) 43.

[13] *Thlimmenos v Greece [GC]* App no 34369/97 (ECtHR, 6 April 2000).

[14] Case C-256/01 *Debra Allonby v Accrington and Rossendale College, Education Lectur-*

possess a different protected characteristic. The people applying for a lorry drivers' licence and a car drivers license are not in similar circumstances and thus it is allowed to have stricter eyesight requirements for a lorry license.[15] The similarity depends on the context and can differ between different member states of the EU. A married couple and an unmarried couple might not be in similar circumstances for the purposes of social security (in one member state), but they could be in similar circumstances with respect to the right to contact family while in custody.[16]

The last element to consider is the causation between the less favourable treatment and the protected ground. If a person would have been treated differently if they held a different protected characteristic, then the treatment is almost certainly discriminatory. An example of this is the case of a gay man whose registered partner died.[17] He wanted to claim the survivors pension offered by the employer of his dead partner, but the company refused to pay since the two men were not married. Here, it is clear than if they would have been allowed to get married, the treatment would be different.

Direct discrimination covers the type of discrimination which is obvious and prohibits it. However, it requires a clear causal relationship between the treatment and the protected characteristic which the victim holds. Often, discrimination is not this direct. Instead of basing treatment on protected grounds, a rule might use characteristics which are related to certain protected characteristics but not a perfect match. This type of discrimination is called indirect discrimination and I

---

ing Services, trading as Protocol Professional and Secretary of State for Education and Employment (CJEU, 13 January 2004).

15 Case C-356/12 *Wolfgang Glatzel v Freistaat Bayern* (CJEU, 22 May 2014).

16 European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (n 1) 47.

17 Case C-267/06 *Tadao Maruko v Versorgungsanstalt der deutschen Bühnen [GC]* (CJEU, 1 April 2008).

will discuss it next.

2.2 Indirect discrimination

Indirect discrimination is more subtle than direct discrimination and occurs when an apparently neutral rule affects a group or person which/who hold(s) a protected characteristic in a significantly more negative way than others in a similar situation.[18] It is a more subtle form of discrimination and harder to detect. The prohibition against indirect discrimination aims to protect the same protected grounds as direct discrimination, but it targets rules which seem neutral, but which have a discriminatory effect. The focus lies on the protection of definable groups of people and how seemingly neutral rules can affect them in very different ways.

The constituent elements of indirect discrimination are different from those which make up direct discrimination. The first element of indirect discrimination is that there must be a rule, criterion or practice which appears to be neutral.[19] For example, rules which disadvantage people who work part-time compared to their full-time peers. At its face, these two groups are so different that they do not have to be treated similarly.

The second element is that the this apparently neutral rule must put a protected group at a particular disadvantage.[20] This means that the group which is disadvantaged must have significantly more people who hold a certain protected attribute. It is not necessary for a group to consist of just people which have the

---

[18] European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (n 1) 54.

[19] European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (n 1) 54.

[20] European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (n 1) 56.

same characteristics, a significant proportion is sufficient.[21] For example, Lufthansa required its pilots to be taller than 1,65 meters.[22] Only 2,6% of men are shorter than 1,65 m but 44,3% of women do not pass this height requirement. Thus, the group of people who were disadvantaged by this treatment where mostly women and Lufthansa ended up paying compensation for unequal treatment.

The last element is the existence of a comparator group.[23] This group must be advantaged by the treatment while the protected group is disadvantaged by it. The comparator group can also contain people with the same protected characteristic as the protected group, but the proportions of the comparator group need to significantly differ. If a rule significantly disadvantages people with part-time contracts, then this is very likely indirect discrimination since women are much more likely to work part-time than men, thus forming a protected group within the larger group of part-time workers. The focus is on the treatment, and it can also affect men and still be indirect discrimination.

The prohibition of indirect discrimination supplements the prohibition of direct discrimination by including discrimination which uses correlation to discriminate. While this extension of the notion of discrimination is needed, it might not be sufficient to deal with discrimination by ADM systems. In Section 2.4, I will introduce a third type of discrimination which might be necessary to define the discrimination which might take place in an ADM system. However, first, I will cover the possibilities for justifying discrimination.

---

[21] European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (n 1) 57.

[22] European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (n 1) 57.

[23] European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (n 1) 57.

2.3 Justifications for less favourable treatment

The rules prohibiting discrimination aim to ensure that every person is treated equally and that everyone gets a fair chance to the opportunities their society offers. However, there are circumstances where both EU and ECHR law accept some differential treatment.[24] Under EU law, there must be an objective justification for indirect discrimination. For direct discrimination, EU law contains three specific exceptions to the prohibition. These relate to requirements for specific occupations, religious organisations and age related discrimination if it passes the proportionality test.[25]

While there are very few exceptions to the prohibition against direct discrimination, EU law provides more room for justifying indirect discrimination. It is possible to justify indirect discrimination based on an objective justification.[26] For an objective justification, the discriminatory rule needs to have a legitimate aim and it needs to be appropriate and necessary for the aim.[27] In order to justify a rule which is indirectly discriminatory, there must be no other appropriate way to achieve the legitimate aim.

---

[24] European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (n 1) 92.

[25] European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (n 1) 97-103.

[26] European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (n 1) 94.

[27] European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (n 1) 94.

2.4 Issues with applying non-discrimination law to ADM systems

In the previous three sections, I have given an overview of the overall EU framework of non-discrimination law. Some authors argue that these two types of discrimination are not broad enough for dealing with ADM systems.[28] They require a categorization of concepts which might not be suited for categorization as well as a relation between a protected attribute and the discrimination which might hard to detect in an ADM system. An example of this would be the treatment of non-binary people in a model which aims to ensure equality between men and women. Any ADM system will require data and that usually means categorization of the real-world simulation, which might be harder than expected for cases like non-binary people.

Another possible issue for dealing with ADM systems is that both the EU and ECHR legal orders lack formal recognition for intersectional and multiple discrimination. The first of these two, intersectional discrimination, occurs when discrimination happens based on a combination of multiple protected characteristics. This is a kind of discrimination which can happen in an ADM system, but there is no explicit legal prohibition against this type of discrimination yet.[29] Multiple discrimination does not require the interaction effect between the protected characteristics but instead occurs when discrimination is aimed at separate characteristics.

Complex kinds of discrimination is not explicitly prohibited in the EU. The lists

---

[28] A Balayn and S Gürses, *Beyond Debiasing: Regulating AI and its inequalities* (2021).

[29] European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (n 1) 59-63.

of protected grounds listed in the equality directives are even exhaustive and cant be extended by case law.[30] Multiple discrimination has been tacitly acknowledged in the case law of the ECtHR. A female sex worker from Nigeria who worked and lived in Spain alleged that the police mistreated her due to her race, gender and profession.[31] She alleged that she was checked and insulted much more than her colleagues of European origin. The ECtHR found that the Spanish courts had failed to take into account the vulnerability of an African woman working as a prostitute. The combination of different attributes and the fact that these lead to a worse treatment than each separately is a form of multiple discrimination.

The last possible issue with applying non-discrimination law to ADM systems is the burden of proof. The person who alleges discrimination needs to provide evidence which suggests that discrimination has taken place.[32] While European non-discrimination law does split the burden of proof between the two parties, where the complainant needs to prove a presumption of discrimination and then it is up to defendant to show that the treatment is not discriminatory. To create a presumption for indirect discrimination, a complainant might want to use statistics. While courts have accepted the use of statistics for this purpose, they do emphasize that these must show a substantial proportion of the group affected by discrimination possesses the protected characteristic.[33] If 60% of the people in a group is female, this might not be enough to establish that the less favourable

---

[30] European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (n 1) 62.

[31] *BS v Spain* App no 47159/08 (ECtHR, 24 July 2012).

[32] European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (n 1) 230.

[33] European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (n 1) 242-243.

treatment of this group is discriminatory based on sex.[34]

2.5 Prohibition of discrimination by ADM systems in the GDPR

The GDPR prohibits ADM systems if its decisions are *based solely* on the automated processing and have *legal or similarly significant effects* on individuals.[35] Decisions are considered to be *based solely* on automated processing if any people involved in the process do not influence the decision.[36] If the people involved have the authority and competence to changed the decision, then the decisions are probably not based solely on the automated processing.

To have a *legal effect*, a decision must affect an individual's legal rights, or their legal status and rights under a contract.[37] The description of what *similarly significant effects* are, is a little less clear than that of legal effects. Examples of what it includes are automatic refusals of credit card applications and online recruiting practices without human intervention.[38] More generally, if an automated decision has the potential to significantly affect the behaviour of individuals, have a lasting or permanent impact on them, or lead to exclusion or discrimination of

---

[34] Opinion C-317/93 *Inge Nolte v Landesversicherungsanstalt Hannover* (CJEU, 31 May 1995).

[35] 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) [2016] OJ L119/1 (GDPR) art 22(1).

[36] Article 29 Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679* (2017) 21.

[37] *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679* (n 36) 21.

[38] GDPR, Recital 71.

individuals, then it has a significant effect.[39]

The GDPR contains three exceptions to the prohibition of ADM. If the decision made by the system is necessary for the performance of a contract, authorised by Union or Member State law or based on the explicit consent then the use of the ADM system is allowed under the conditions set out in Article 22.[40]

When using an ADM system, the data controller must implement suitable measures to protect the fundamental rights and freedoms of the data subject and their legitimate interests.[41] While Article 22 GDPR does not specify what these measures are, Recital 71 GDPR contains a list of actions which the controller of an ADM system should take. The controller should use appropriate statistical procedures in developing their ADM system, implement measures to ensure that the risk of errors is minimised, and secure personal data in order to mitigate potential risks for the rights and interests of data subjects and discriminatory effects based on protected grounds. Thus, any controller which uses an ADM system must ensure that their system does not discriminate either directly or indirectly and take measures to ensure this.

---

[39] *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679* (n 36) 21.

[40] GDPR, art 22(2).

[41] GDPR, arts 22(2)(b) and 22(3).

3 Measuring discrimination

One of the characteristics of an ADM system is that its outcomes are automatically generated. Since the decisions of an Article 22 GDPR are not systematically checked by a person capable of changing them, if the results are discriminatory then the ADM system will discriminate automatically. In this section, discrimination can be either direct, indirect or even multiple in nature. The legality of the discrimination will not be analysed yet.

In the computer science literature, discrimination in a system is called bias.[1] A system without any bias is considered to be fair. Since an ADM system is intended to replace the work of (many) people, an automated way to search for bias is necessary to allow for effective oversight, both during the development of the ADM system and during its deployment. The checks during development allow the developer to take measures to mitigate the discrimination while the checks after development allow the user of the ADM system to detect the discrimination and take action to mitigate it.

3.1 Measuring fairness

The objective of non-discrimination law is that every member of a society is treated the same way under similar circumstances and that every member of society has an

[1] A Balayn and S Gürses, *Beyond Debiasing: Regulating AI and its inequalities* (2021).

equally fair chance of taking the opportunities which their society offers.European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (2018 edition, Publications Office of the European Union 2018) In order for a measure of fairness to detect direct or indirect discrimination, fairness metrics should be able to compare outcomes between groups with different protected characteristics who are in similar circumstances. In this subsection, I will describe a number of possible ways to measure fairness, selected from a literature summary of metrics.[2] All of them are meant for yes-no outcomes, so called classification systems.

3.1.1 Notation

Since these measures all aim to summarise the level of fairness in a single number, some notation will be necessary to define the measures. While many readers will not be familiar with this type of mathematical writing, I have included it for completeness sake in addition to a textual description of the metric. I will use the following notation to describe the fairness measures:

- S_i represents the protected attribute of an individual i.
- X_i represents all the other attributes of an individual i.
- Y_i represents the actual outcome label of an individual i.
- $Pred(c)_i = P(Y = c|S_i, X_i)$ represents the predicted probability of a certain outcome c by the ADM system based on S_i and X_i.
- d_i represents the predicted outcome for an individual i. If $Pred(c)_i$ is larger than a certain threshold, then $d_i = c$.

[2] S Verma and J Rubin, 'Fairness definitions explained' (FairWare '18, Association for Computing Machinery May 2018).

3.1.2 Metrics

I have selected three metrics based on their intuitiveness and their ease of implementation. The first metric is a very simple conception of fairness which only looks at the outcomes at a group level. The other two metrics look at an individual level at the differences in outcome when you take the attributes of an individual into account.

Group fairness:[3] Fairness according to this metric is if individuals with different S values have an equal probability of getting a positive result, $P(d = 1|S = a) = P(d = 1|S = b)$. This metric represents the idea that there should be no relation between S and the outcome of an ADM system on a group level.

Causal discrimination:[4] Fairness according to this metric is if individuals x and y with either value of S and the same attributes X get the same predicted outcome, $(X_x = X_y \wedge S_x \neq S_y) \rightarrow d_x = d_y$. This metric emphasises that people with similar attributes should be treated the same.

Fairness through unawareness:[5] Fairness according to this metric is if individuals x and y with similar values of X get the same predicted outcome, without looking at S at all, $X_x = X_y \rightarrow d_x = d_y$. This metric emphasises that people with similar attributes should be treated the same, like with Causal discrimination, but that protected attributes should just not be used.

---

[3] C Dwork and others, 'Fairness Through Awareness' [2011] (arXiv:1104.3913 [cs] type: article).

[4] S Galhotra, Y Brun, and A Meliou, 'Fairness testing: testing software for discrimination' (ACM August 2017).

[5] MJ Kusner and others, 'Counterfactual Fairness' [2018] .

3.2 Case study

In order to test the effectiveness of the technical methods of the next chapter, I constructed an artificial dataset which will serve as a case study of a discriminatory dataset. The discrimination model which I will use is inspired by the work of Kamiran,[6] in that it assumes three different steps in the decision-making process. The dataset will be made-up of prospective students applying to a fictional Dutch university for a study which will be taught in English and their admittance evaluation. The dataset will be used to train an ADM system which has to decide which students should be admitted. Since the result of university admittance 'significantly affects' affects a person in a similar way to a legal effect, this type of system can be an example of automated decision making according to the Article 29 Working Party.[7]

3.2.1 Discrimination model

The (fictional) selection process of the university works with three steps. All students make a placement-test and receive a certain score. Based on this score and other relevant factors, a selection committee gives the students a rating, but this rating is biased towards Dutch students. Lastly, all students with a rating above a certain level are admitted into the study. In Table 3.1, you can see the exact details of each variable.

We know how the decision is made since the discrimination model is known to us. In the sections on measuring and removing discrimination, I will assume that

[6] F Kamiran, I Žliobaitė, and T Calders, 'Quantifying explainable discrimination and removing illegal discrimination in automated decision making' (2013) 35(3) Knowledge and Information Systems 613.

[7] Article 29 Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679* (2017).

Variable	Description	Generating Mechanism
Nationality	Nationality of the Applicant, Dutch or Non-Dutch	Equal chances for Dutch and Non-Dutch
Gender	Gender of Applicant, Male or Female	Equal chances for Male and Female
Test	Score on placement test, from 1 to 10	Normal distribution Mean = 6.5, SD = 1.5 Rounded to 1 digit
English-Certificate	Applicant has an English certificate, Yes or No	Dutch: 10% of the time Yes Non-Dutch: 60% of the time Yes
Extracurricular	Applicant has extracurricular activities, Yes or No	Dutch: 20% of the time Yes Non-Dutch: 60% of the time Yes
Rating	Rating by the Committee, from 1 to 10	Dutch: Test + 0.8 (Still max 10) Non-Dutch: Test
Accepted	Accepted for study, Yes or No	If Rating $\geq$ threshold, then Yes, else, No

Table 3.1: Description of the case study dataset and its generating mechanism.

the person making the ADM system is not aware of the *Rating* attribute. That person will also not know that *English Certificate* and *Extracurricular* have no impact on the acceptance (since these are ignored in the discrimination model).

This model for discrimination ensures that the privileged group (Dutch applicants) have a higher chance of being accepted. The applicants which are disadvantaged the most by this are the non-Dutch applicants who scored a little below the threshold, but who do not have the advantage of getting bonus from the committee. In the paper which inspired this discrimination model, the authors highlight that the addition of a bias factor (via *Rating*) corresponds with empirical findings about the effect of discrimination on chances and decisions.[8]

[8] Kamiran, Žliobaitė, and Calders, 'Quantifying explainable discrimination and removing illegal discrimination in automated decision making' (n 6).

Variable	**Population n=1000**	**Accepted n=163**	**Not Accepted n=837**
Nationality	49.7% Dutch	65.6% Dutch	46.6% Dutch
Gender	51.4% Female	47.2% Female	52.2% Female
Test	Mean = 6.51 SD = 1.49	Mean = 8.68 SD = 0.63	Mean = 6.09 SD = 1.21
English Certificate	33.4% Yes	28.2% Yes	34.4% Yes
Extracurricular	40% Yes	37.4% Yes	40.5% Yes
Rating	Mean = 6.90 SD = 1.50	Mean = 9.18 SD = 0.50	Mean = 6.46 SD = 1.20

Table 3.2: Summary statistics of the training dataset.

3.2.2 Generating the data

To generate the data based on the discrimination model from the previous subsection, I implemented the model in an R script (R Core Team 2021). All the code can be found in the GitHub repository.[9] First of all, I generated a population of 1000 applicants based on the characteristics I described in the previous Subsection. Then, I applied the rating filter and decided on a threshold of 8.5 for automatic acceptance. Table 3.2 shows some general characteristics of the overall population and the two subgroups. You can see that, as expected, the rating discrimination has resulted in a group of accepted students which consists of 107 Dutch applicants and 56 people with other nationalities. This distribution is very different from the 50-50 of the population of applicants. This is an indication that there is a form of discrimination going on.

3.2.3 Applying the metrics to the generated data

To see how the three metrics which I defined at the start of this chapter work, I applied them all to the data generated by the discrimination model. I will discuss

[9] https://github.com/plofknaapje/Thesis-ADM-Discrimination

the metrics in the same order as I introduced them.

First of all, the generated outcomes are not fair according to group fairness. I trained a decision tree[10] on all the attributes except *Rating* to predict the value of *Accepted*.[11] It can model the decision process with 100% accuracy. The chance of being accepted is 16.3% for the entire population. Of the 497 Dutch applicants, 107 were accepted, which is 21.5%. Compare that to the Non-Dutch applicants, where out of 503 applicants, just 56 were accepted, or 11.1%. For group fairness, these two probabilities should be equal and since this is not the case, the generated outcomes are not fair.

To measure the causal discrimination of the generated outcomes, I trained a decision tree just like for the group fairness. This allowed me to generate predictions for counterexamples for all the applicants. Out of the 1000 applicants, 123 of them got another outcome if their *Nationality* was flipped. Thus, the generated data causes causal discrimination in the decision tree.

To investigate the effectiveness of fairness through unawareness on the generated data, I trained a decision tree again, this time only on the *Test*, *English Certificate* and *Extracurricular* variables. The decision tree accepts those applicants with *Test* higher than 8.45 or *Test* higher than 7.65, *English Certificate* = No, and *Extracurricular* = No. This is not directly discriminatory, since there are Non-Dutch applicants who satisfy these criteria. However, this model still results in more Dutch applicants being accepted, 95 in total, than Non-Dutch applicants, 68. Thus, the predictions based on the generated data still discriminatory under unawareness of the sensitive attribute.

---

[10] A decision tree model is based on a sequence of decisions, like *Test* ¿ 8.5 or *Nationality* = Dutch.

[11] G James and others, *An introduction to statistical learning: with applications in R* (Second edition, Springer texts in statistics, Springer 2021).

These three results show that there is discrimination present in the generated data. The last result also shows that just removing the sensitive attribute is not sufficient to remove the discrimination.

3.3 Limitations of this approach

Before I continue to the next chapter, I want to highlight the limitations of the approach I have laid out above. With respect to the fairness measures, there are authors who are critical about the measurement-based approach to bias.[12] They consider this approach to be too limited since the bias of an ADM system is checked with an existing dataset, which has been looked at during the development process. This can lead to problems since during its use the data the ADM system will be processing will most likely differ in many ways from the data it was tested on. Others argue that it is impossible to automate anything like fairness since courts consider discrimination to be contextual, which would require fairness metrics to incorporate the context for its evaluation which makes it harder to achieve the goal of automated monitoring for bias.[13]

With regards to the discrimination model, it is a way to generate a simplified version of an actual dataset which assumes that there is a good way to grade a job application in an automated way. Further assumptions are that men and women are equally qualified and apply in equal amounts to this job. These assumptions lead to a dataset where at least men and women are equally represented but with large differences in the effects for these two groups. The model also only considers men or women, with no regard for those who do not conform to this

[12] Balayn and Gürses (n 1).

[13] S Wachter, B Mittelstadt, and C Russell, 'Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI' [2020] (ID 3547922) .

binary classification. This is also a problem with the fairness measures, since they require a way to define the different groups within a protected attribute, but for many protected attributes, it might not be possible to do so in a way which respects everyone's identity.[14]

[14] Balayn and Gürses (n 1).

4 Technical methods to remove discrimination

In the previous chapter, I described ways to measure the fairness of an ADM system. I also constructed a biased dataset to test technical methods which remove bias from ADM systems. In this chapter, I will give an overview of the technical methods which aim to remove discrimination from an ADM system. Technical methods can be applied at three different stages of the development of an ADM system.[1] They can be used before the training of the ADM system (pre-processing),[2] during the training of the ADM system (in-processing) or after the ADM system has been trained (post-processing). I will discuss each of these types of methods in their own section. For each specific method in the sections, I will describe how the method works and also explain why its approach is fair to apply.

The three types of technical methods respond to different sources of bias in an ADM system. If the training data is high quality and suitable for the goal of the ADM system, then better data will lead to more accurate predictions.[3] A system is accurate if its predictions match with the value which has to be predicted. The

---

1 N Mehrabi and others, 'A Survey on Bias and Fairness in Machine Learning' [2022] .

2 A Balayn and S Gürses, *Beyond Debiasing: Regulating AI and its inequalities* (2021).

3 A Agrawal, J Gans, and A Goldfarb, *Prediction machines: the simple economics of artificial intelligence* (Harvard Business Review Press 2018).

value to be predicted is biased, like in the case study, then being accurate is not the same as being fair. In practice, it might not be possible to obtain data which is complete and represents the entire population or which is completely accurate.[4] The training data of the system can be biased in many different ways.[5] The ADM system can also turn out to contain biased parameters after its training is done.

All the methods in this chapter aim to perform some kind of intervention which makes the treatment of people depend less on the sensitive attribute S. They consider any difference in treatment to be undesirable, which fits with the structure of the case study from Chapter 3. However, there might be situations where there is a good explanation for a difference in treatment between two groups. An example would be a university where more women apply for study A, which has 50 spots, and more men apply for study B, which has 100 spots.[6] If there is no distinction made on a study level, the university ends up with more male students overall, which might look discriminatory, but which can be explained and justified.

4.1 Pre-processing methods

Technical methods in the pre-processing stage aim to change the training data and remove their bias before the ADM system gets a chance to internalise it.[7] After these pre-processing methods, the development of the ADM system can then

---

4 European Union Agency for Fundamental Rights, *Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights* (2019).

5 Mehrabi and others (n 1).

6 F Kamiran, I Žliobaitė, and T Calders, 'Quantifying explainable discrimination and removing illegal discrimination in automated decision making' (2013) 35(3) Knowledge and Information Systems 613.

7 F Kamiran and T Calders, 'Data preprocessing techniques for classification without discrimination' (2012) 33(1) Knowledge and Information Systems 1.

(hopefully) be done with regular algorithms. The methods described below all use the sensitive attribute S during the pre-processing, to change the dataset such that S is not necessary later to remove discrimination or for training the model. The three proposed techniques are called massaging, reweighting and sampling.[8] They all aim to reduce the discrimination in the dataset to 0, which is defined in this paper as the difference between the probabilities of being in the positive class between two groups, which is similar to the concept of group fairness from Section 3.1. I will illustrate these methods with some example data (Table 4.1).

These three methods all assume either that there is no relation between S and the outcome label Y, or that if there is a relation that it should not be included. Thus, they all aim to change the dataset in such a way that the relation between S and Y disappears. This modification makes the dataset a bit artificial, since information is removed to achieve equal outcomes, but if you accept that the circumstances under which the data was collected lead to discriminatory outcomes, then changing the dataset to be less discriminatory could be a good thing. It is important to note that the people in the dataset who benefited from the discrimination will not be disadvantaged by these methods, since only new decisions will be made by this. As for the people in the privileged group who complain that these measures harm their future chances,[9] I say: when you are used to privilege, equality feels like oppression.

Massaging: The aim of the massaging technique is to change the value of Y from 0 to 1 for some individuals in the dataset and from 1 to 0 for an equal number of individuals. The people who will be swapped are determined based on

---

[8] Kamiran and Calders, 'Data preprocessing techniques for classification without discrimination' (n 7).

[9] Why do people resist EDI initiatives?, 'Association of Medical Research Charities' <https://www.amrc.org.uk/blog/why-do-people-resist-edi-initiatives> accessed 17 July 2022.

an algorithm which aims to predict Y. The people of the discriminated group who have $Y = 0$ but a very high ranking have their value of Y flipped and the reverse happens to the people from the privileged group.

The two groups of people, Dutch and Non-Dutch, have the same values for Score and Experience but different outcomes in Table 4.1. Here, person 3 is the least qualified Dutch person, with a lower Score and lower Experience than person 6. Thus, when massaging the dataset, person 3 would be flipped from 1 to 0 and person 6 would be flipped from 0 to 1. Thus, the total number of positive outcomes stays the same but they are now equally distributed over the two values of *Nationality*.

Reweighting: This technique is less aggressive than massaging, since it does not change the labels of the dataset. Instead, reweighting aims to make instances which are beneficial for the disadvantaged group more impactful on the outcome of an algorithm. Each individual instance in the dataset gets a weight based on the expected probability of seeing such an instance with its S and Y values.

The weights of the different instances are adjusted based on the chance of seeing that outcome if the positive outcomes were not biased. A weight of 1 is standard, less than 1 leads to a lower impact of that instance and larger that 1 lead to a bigger impact than normal. For Table 4.1, the instances 1, 2, 3, 6, 7 and 8 will get a weight of less than 1 and on the other hand the two 'rare' outcomes, 4 and 5, will get a weight of more than 1.

Sampling: Since not all learning algorithms are capable of dealing with weighted datasets, an alternative is to sample a new dataset based on those weights. For this, the dataset is divided into four groups, based on the values of S and Y. For each group, the expected size of the group is determined. With this information, the authors propose two different approaches: Uniform sampling and preferential sampling. I will use uniform sampling, where every point in each

Index	*Nationality*	*Score*	*Experience*	*Outcome*
1	Dutch	5	3	1
2	Dutch	4	1	1
3	Dutch	3	2	1
4	Dutch	2	0	0
5	Non-Dutch	5	3	1
6	Non-Dutch	4	1	0
7	Non-Dutch	3	2	0
8	Non-Dutch	2	0	0

Table 4.1: Eight possible applicants as an example for explaining the methods.

group has an equal chance of being either duplicated or removed in sampling. In preferential sampling, the points closest to the decision boundary (the cut-off which decides which *Rating* is needed for $d = 1$) are changed first.

The sampling method divides the data into four groups of Table 4.1 based on the *Nationality* and *Outcome* variables. Since these groups should all contain two people if the positive outcomes were equally divided based on *Nationality*, the sampling method will pick 2 of the 3 instances in the groups which contain 3 people and use the same person twice in the groups which consist of just 1 person. This leads to a new dataset where the positive outcomes are spread equally over the two *Nationality* values.

4.2 In-processing methods

methods applied during the training of an ADM system aim to change the objective of the system or the way it learns in order to prevent discriminatory outcomes. These methods all aim to adjust the way existing algorithms learn in order to strike a balance between the accuracy of the resulting algorithm and the fairness of the outcomes.

Just like the pre-processing methods from the previous subsection, these meth-

ods all assume that there is either no relation between S and Y, or that this relation should not be included. Just like with the pre-processing methods, if you accept that the training dataset includes a measure of discrimination and if you are against discrimination, then trying to remove discrimination seems sensible. These measures focus the relevant algorithm to focus more on other aspects of the dataset and also to counteract the discrimination which is present in the dataset. In contrast to the pre-processing methods, these methods require S when making new predictions, which is a major disadvantage.

Modifying naive Bayes: This method takes advantage of the output of the naive Bayes model.[10] This classifier calculates a probability score for each possible outcome. By changing the probability score required for a positive prediction for the discriminated or privileged group, it is possible to disrupt the relation between S and Y. This allows the removal of discrimination from the outcomes of the model.[11]

While the first three methods adjusted the dataset, this method adjusts a default model. After training a naive Bayes classifier, the relation between S and Y is adjusted to make sure that the difference between the two group fairness scores is as low as possible. There is no guarantee that it is possible to achieve total equal outcomes with this.

Classification under Fairness: This method is capable of focussing on either fairness or accuracy while taking the other into account. The method looks at the decision boundary of the learning algorithm and aims to adjust this in order

---

[10] The naive Bayes model learns the correlations between the different values of variables and the outcome and uses this to determine which outcome is the most likely; (G James and others, *An introduction to statistical learning: with applications in R* (Second edition, Springer texts in statistics, Springer 2021) 37).

[11] T Calders and S Verwer, 'Three naive Bayes approaches for discrimination-free classification' (2010) 21(2) Data Mining and Knowledge Discovery 277.

to increase the fairness of the outcomes of the algorithm. Due to the way it is designed, this is the only method designed to deal with sensitive attributes with more than two values and datasets which contain multiple sensitive attributes.[12]

This method is the most complicated of the six. It allows you to determine the importance of fairness and accuracy respectively. Focussing on accuracy will lead to similar outcomes as the decision tree model from Subsection 3.2.3. Focussing on fairness will lead to the model with the least discrimination.

4.3 Post-processing methods

This last type of methods accept that the underlying models might not be able to be adjusted in a way which removes discrimination. Thus, they try to adjust the outcomes of models which have already been trained but which are still discriminatory. This means that, just like with the in-processing methods above, S is required to make new predictions. This is a major disadvantage compared to the pre-processing methods.

Naive Bayes ensemble: This method trains two different naive Bayes models, one for each value of S. These both aim to predict Y. This results in two models which try to predict Y and do so with a certain probability $Pred(c)$. These two models can then be balanced to achieve overall outcomes which do not discriminate by adjusting the thresholds for a positive outcome for the two groups of applicants.[13]

With this method, it is accepted that the two different groups do not get the same prediction results. So, now there are two different models and based on your nationality, applicants get assigned to one of the two. Then, by looking at the spread of the outcomes, the results can be balanced by accepting people in from

[12] MB Zafar and others, 'Fairness Constraints: Mechanisms for Fair Classification' [2017] .

[13] Calders and Verwer (n 11).

the disadvantaged group more. This lower bar for acceptance can compensate the discrimination in the dataset.

4.4 Limitations

All these methods do require access to the sensitive attribute S, at least initially. Some of them aim to remove the necessity of this attribute by balancing the training data in a way to remove the effect of proxy variables. All of the methods have a shared principle: the outcome Y should not depend on S. This is in line with non-discrimination law. To make sure that Y no longer depends on S, the methods have multiple possible approaches. They can change the data in such a way that it no longer includes discriminatory relations between S and Y or they can come up with different rules for people with different values of S. While the methods are quite drastic in their approaches, the effects of a discriminating ADM system make it worthwhile to at least examine the effects of anti-discrimination methods and see how they work.

5 Evaluation of the technical methods

In this chapter, I will combine the metrics from Chapter 3 and the methods from Chapter 4 by applying them to the dataset generated in Chapter 3. I will evaluate the methods based on their scores on the three metrics and the extent to which a method requires access to the protected attributes.

5.1 Applying the methods to the dataset

I used the programming languages R[1] and Python[2] to implement the six methods. Below, I will give more details for each method. The results of the tests of each method can be found in Table 5.2. All these results can also be found in the GitHub repository. For each notebook in the notebooks folder, there is a pdf version which shows the results of running the code without requiring the viewer to have an R environment. All the metrics were computed by applying the models to a secondary test dataset generated based on the discrimination model of Chapter 3. Table 5.1

---

[1] R Core Team, R: A language and environment for statistical computing (2022) <https://www.R-project.org/>.

[2] Python Software Foundation, The Python Language Reference (Version 3.10.5,) <https://docs.python.org/3/reference/> accessed 7 July 2022.

Variable	Population (n=1000)
Nationality (sensitive attribute)	50.4% Dutch
Gender	48.7% Female
Test	Mean = 6.50 SD = 1.49
English Certificate	33.6% Yes
Extracurricular	40.1% Yes
Rating	Mean = 6.88 SD = 1.53

Table 5.1: Summary statistics for the test dataset.

shows the statistics for this test dataset.

5.1.1 Implementation details

In this subsection, I will give a brief description of how I implemented each method. Most of them were implemented in R. Classification under Fairness is the exception, since it was originally developed in Python. For the Reweighting and Resampling methods, I was able to use an existing package, the fairmodels package.[3] For the other three methods, I implemented them based on their description in their respective papers. None: The None method is the default situation of the dataset with standard models. These outcomes are the same as the results from Chapter 3.

Massaging: I implemented the massaging procedure based on the paper.[4] I trained a naive Bayes model and made a prediction for each applicant. The Dutch applicants who got accepted and had very low predictions were flipped to not accepted. The Non-Dutch applicants who were not accepted and got high predictions were flipped to accepted. This procedure was done until both groups

[3] J Wiśniewski and P Biecek, 'fairmodels: A Flexible Tool For Bias Detection, Visualization, And Mitigation' [2022] .

[4] F Kamiran and T Calders, 'Data preprocessing techniques for classification without discrimination' (2012) 33(1) Knowledge and Information Systems 1.

had an equal percentage of accepted people. Then, I trained a decision tree model on the massaged dataset and computed the three metrics.

Reweighting: I used the reweight function from the fairmodels package.[5] It implements the reweighting procedure from the paper. With their reweight function, I trained a weighted naive Bayes model and then applied the metrics to its predictions.

Resampling: I used the resample function, set to uniform, from the fairmodels package.[6] It implements the resampling procedure from the paper. Then, I trained a decision tree model on the resampled dataset and applied the metrics to its predictions.

Modified naive Bayes: I implemented the modification procedure based on the paper.[7] First, I trained a naive Bayes model and made a prediction for each applicant. The normal cut-off to get accepted is 0.5 and the model makes predictions on a range from 0.0 to 1.0. Now, I made a separate cut-off for Dutch and Non-Dutch applicants. Then, I incrementally lowered the cut-off for Non-Dutch and increased it for Dutch people in such a way that the number of accepted people stayed the same, but that the proportions of Dutch and Non-Dutch became the same.

Classification under Fairness: I used the Python code of Muhammad Bilal Zafar, one of the authors of the paper which proposed this method.[8] I imported the dataset in such a way that it fit with the system which Zafar designed. Then, I trained a standard version of their model and a fairness focussed one. The fairness

---

5 Wiśniewski and Biecek (n 3).

6 Wiśniewski and Biecek (n 3).

7 T Calders and S Verwer, 'Three naive Bayes approaches for discrimination-free classification' (2010) 21(2) Data Mining and Knowledge Discovery 277.

8 MB Zafar and others, 'Fairness Constraints: Mechanisms for Fair Classification' [2017] .

through unawareness result is from the standard version of their model, which is unaware of the sensitive attributes.

Naive Bayes ensemble: I implemented this procedure based on the paper.[9] This method requires separate datasets with all the applicants with one sensitive attribute, so I made a Dutch dataset and a Non-Dutch dataset. Then, for each group I trained a separate model. The method for removing the bias is the same as for Modified naive Bayes. I made a separate cut-off for each model. Then, I lowered the Non-Dutch cut-off and increased the Dutch cut-off incrementally to obtain an equal percentage of acceptance while keeping the number of accepted applicants the same.

5.1.2 Results

In Table 5.2, you can see the outcomes of the metrics for each method. For some methods, Fairness through Unawareness does not make sense. This is indicated with the 'Not Applicable'. These methods require the use of the protected attribute during predictions, which means that the prediction can never be unaware. In this subsection, I will interpret the results for each method in combination with the way in which the result was reached.

None: These results are similar to the results from Chapter 3. A model which takes nationality into account will discriminate against Non-Dutch applicants and when nationality is not included in a model, the outcomes are still biased. There is also a high amount of causal discrimination, which indicates that nationality plays a big part in the model.

Massaging: The massaged training dataset delivers good results, as long as nationality is included. Dutch and Non-Dutch applicants have an equal chance of

[9] Calders and Verwer (n 7).

Method	**Group Fairness** % Accepted (Total Accepted)	**Causal Discrimination** % with different outcome if nationality is changed	**Fairness through Unawareness** % Accepted (Total Accepted)
None	Dutch: 21.5% Non-Dutch: 11.1% (161)	13%	Dutch: 20.4% Non-Dutch: 10.5% (163)
Massaging	Dutch: 16.3% Non-Dutch: 16.3% (163)	0%	Dutch: 19.2% Non-Dutch: 15.5% (174)
Reweighting	Dutch: 24.8% Non-Dutch: 8.7% (168)	13%	Dutch: 24.8% Non-Dutch: 21.4% (231)
Resampling (uniform)	Dutch: 16.3% Non-Dutch: 16.3% (163)	13%	Dutch: 20.4% Non-Dutch: 10.5% (155)
Modified naive Bayes	Dutch: 15.7% Non-Dutch: 17.5% (168)	1.2%	Not Applicable
Classification under Fairness	Dutch: 17.5% Non-Dutch: 12.7% (150)	0%	Dutch: 17.7% Non-Dutch: 12.3% (151)
Naive Bayes ensemble	Dutch: 19.8% Non-Dutch: 12.7% (163)	7.1%	Not Applicable

Table 5.2: Results of the metrics for each method.

being accepted and there is no causal discrimination. If nationality is not included, then the massaged dataset loses its effectiveness. The results are still better than None, but not free from bias.

Reweighting: Reweighting performs poorly. The results are even worse than those of *None* when computing group fairness and the same for causal discrimination. Under unawareness, the selection is not strict enough and too many people get accepted. The gap between the Dutch and Non-Dutch acceptance rates does decrease considerably under unawareness.

Resampling: The results of resampling are a bit strange. On group fairness, its scores well. However, the causal discrimination and the unawareness results are the same as those of *None*. Thus, the resampling model does require the inclusion of nationality.

Modified naive Bayes: This method is a bit too aggressive. It overcompensates for the bias against Non-Dutch applicants. In the end, the cut-off for Dutch applicants stayed at 0.5 and the Non-Dutch cut-off got lowered to 0.35. The model exhibits almost no causal discrimination, which is good.

Classification under Fairness: This method performed well based on its own evaluation when just using the training data, but when tested on the test dataset, the results were much worse. It managed to remove some of the bias but not in a meaningful way. Its one redeeming feature is the lack of causal discrimination.

Naive Bayes ensemble: This method performed poorly. Even with a cut-off of 0.7 for Dutch applicants and 0.02 for Non-Dutch applicants, there is still a big difference between the acceptance rates of the two groups according to group fairness. There is also a considerable amount of causal discrimination. This result might be due to the fact that this dataset violates one of the assumptions of naive Bayes, which is that the attributes are not correlated. This is the case in this dataset.

		Sensitive data during pre-processing or training	
		No	Yes
Sensitive data during prediction	No	Fairness under Unawareness: Classification under Fairness.	Fairness under Unawareness: Massaging, Reweighting and Resampling Group Fairness: Classification under Fairness
	Yes	-	Group Fairness: Massaging, Reweighting, Resampling, Modified Naive Bayes, Naive Bayes ensemble (and None).

Table 5.3: Overview of usage categories

5.2 Evaluation of the methods

In the previous Section, I laid out the testing results of the six methods on the three metrics. For the legal evaluation, I will split these combinations up into three different categories, based on whether sensitive data is used during pre-processing or training and during prediction. If sensitive data is used during prediction, it also must be present during training, so one of the fields in Table 5.3 will remain empty. The methods are classified in combination with both group fairness and fairness under unawareness, which means that most of them appear in two categories.

The three different categories require different amounts of the processing of sensitive personal data. This is relevant since, as I highlighted in Chapter 1, the GDPR does not consider the prevention of discrimination a valid ground for processing sensitive personal data.[10] In combination with the data minimization principle, if a method which requires less or even no access to sensitive personal data, it should be preferred to a more invasive method.

[10] 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) [2016] OJ L119/1 (GDPR) art 9.

5.2.1 No-No: no usage of nationality

There is one method which can function while not using nationality during either training or prediction, namely Classification under Fairness. This method scores better than *None*, but it does not perform well enough to consider its results useful.

5.2.2 Yes-No: nationality used during pre-processing or training

The three pre-processing methods can all be used in such a way that the sensitive data is not needed during prediction. This means that if the training dataset could be anonymised, these methods could be used without processing any sensitive data. Unfortunately, the results of these methods on fairness through unawareness are not very good. Massaging has a larger difference between the Dutch and Non-Dutch acceptance rate than Reweighting (which is not selective enough). Resampling performs just as poorly as *None*. Classification under Fairness can be trained with the sensitive data, but does not require the sensitive data for predictions. However, it has a larger difference between the two groups than Massaging.

5.2.3 Yes-Yes: nationality used during training and prediction

The methods in the last category require the sensitive data for both training and prediction. This will be difficult to combine with lack of a valid ground for processing that sensitive data. Of all the methods, Massaging performs the best. The acceptance rates are the same for both Dutch and Non-Dutch applicants and there is no causal discrimination.

5.3 Limitations

The effectiveness of the methods which I tested in this chapter varied a lot. Some were able to achieve totally equal results while other did not remove any bias. The only method which was able to remove all of the bias and leave no causal discrimination was Massaging with the sensitive attribute in the prediction. This shows that it is, unfortunately, necessary to know sensitive attribute in order to make unbiased predictions. This means that the sensitive data must be collected and processed in order to make new predictions. This is an additional type of data processing which might not be allowed by the GDPR.

To come back to a point which I already brought up in Section 2.4, the methods which use the sensitive data require it to be structured. Some of the methods can only deal with two categories and the others still require a categorisation of the different groups. This means that the organisation using these methods to remove discrimination must choose and enforce this categorisation. It also means that if a person does not specify a certain attribute, maybe because they do not identify with any of the options, then the methods which require the sensitive data during prediction might not be able to make a prediction. It might also happen that an empty attribute ruins the prediction.

6 Conclusion and limitations

This chapter combines the main findings of the four preceding chapters, in order to answer the four sub questions from Chapter 1.1. Each question deals with a different topic: non-discrimination law, metrics, methods and results. By combining these summaries, an answer to the main research question will be given. The question was 'Are there useful technical methods to mitigate illegal discrimination in automated decision-making systems?'

6.1 Non-discrimination law

Chapter 2 discusses the European framework for non-discrimination law. Discrimination on protected attributes like race and sex is prohibited in most public interactions. The prohibition of discrimination is also included in the GDPR and explicitly also for systems where decisions which have large effects on people are made based solely on automated processing.[1]

In order to determine if a system is discriminating, it is necessary to know based on what it could be discriminating. A problem with this is that this information is

[1] 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) [2016] OJ L119/1 (GDPR) arts 22 and Recital 71.

often sensitive data.[2] Article 9 GDPR does not allow for processing sensitive data in order to remove discrimination. Thus, not using the data is the best option, but if the ADM system does not use the sensitive data while predicting, then it might be possible to use an anonymised dataset to make the ADM system.

6.2 Metrics

The three metrics discusses in Chapter 3 are used to determine if a method can remove bias. Each metric focusses on different aspects of the concept of discrimination. Group fairness looks at the outcomes for the different groups within a protected attribute. Causal discrimination focusses on the effect of the protected attribute on the outcome in a specific model. Fairness under unawareness shows the effect of not using the protected attribute when making predictions. The combination of these three gives an idea of what a method is capable of and they can be measured automatically. They are unfortunately not able to distinguish between illegal and justified discrimination, since the justification for that is always a human task.

This chapter also includes the case study with which the methods will be evaluated. The case study was designed to contain just one source of discrimination, but with multiple correlating factors. The discrimination in it is very clear as long as you believe that the two groups are equally capable. This case is quite simple, so it forms a baseline for any method. The most effective method should be tested with more complicated data to test how it performs under more complicated circumstances.

[2] GDPR, art 9.

6.3 Methods

Six different methods to produce models which do not discriminate were introduced in Chapter 4. These can be divided into three categories. The pre-processing methods aim to adjust the dataset before the training of the model in order to counter-act the discrimination inherent to the dataset. The in-processing methods adjust the model after training it on the biased dataset. The post-processing method tries to balance the results of two models after training them. All of the methods can theoretically produce models which do not discriminate. They can however not distinguish between justified and illegal discrimination, just like the metrics from the previous section.

6.4 Effectiveness of the methods

In Chapter 5, the six methods were tested. They were tested with a test dataset and evaluated using the three metrics. The most effective method of the six was massaging, which is a pre-processing method. This method changes the training dataset in order to remove its bias. Then, any model can be trained on it. In this case, this was a decision tree. Massaging resulted in no causal discrimination and equal group fairness values. The other methods either result in causal discrimination or their group fairness scores are not the same.

The massaging method is easy to implement and can also possibly be extended to deal with attributes with more than two possible values. It does require the user of the ADM system to accept the fact that the training data is modified. This might not be acceptable to some people, since it goes against the idea of using data to train models.

6.5 Limitations

Massaging requires the use of the protected attribute, so if it is not possible to use the protected attribute at all in the development process due to Article 9 GDPR, then this creates a problem with removing bias. The three metrics also depend on the usage of the protected attribute to detect discrimination. Since the GDPR does require users of ADM systems to ensure that their systems do not discriminate, this could maybe be a justified use of the protected attribute, even if it is sensitive data.

Next to the Article 9 problem, the metrics still cannot distinguish between illegal and justified discrimination, so for that there will always need to be a person to decide what should or should not be removed. While this involvement of people might seem troublesome to the users of ADM systems, Article 22 already requires them to allow data subjects to appeal the decision of the system to a person. For data subjects, the situation where a person must decide if the method is working (too) well is more likely to result in a fair system than the (illegal) situation where an ADM system is implemented without any safeguards.

There are also more practical limitations. First, the technical approach of removing discrimination does require knowledge about the protected attributes of the people involved, while they might not be willing to give this information. This might be a big problem with implementing these methods in practice. Second, my investigation was done on a small dataset with a clear model for discrimination which was applied in a consistent manner.

6.6 Conclusion

Are there useful methods for removing illegal discrimination? Massaging looks like a good candidate, better than the five others. It is capable of removing the bias in the case study, and does so in a transparent way. It can also be combined with any type of learning algorithm. Massaging is however not capable of determining what amount of that bias is illegal discrimination. This remains an open problem and possibly an unsolvable problem. A method like massaging will almost certainly not be able to remove all the bias in a real world application, but it might be a useful tool in the development of fair ADM systems.

Bibliography

Books

Agrawal A, Gans J, and Goldfarb A, *Prediction machines: the simple economics of artificial intelligence* (Harvard Business Review Press 2018).

European Union Agency for Fundamental Rights, *Handbook on European data protection law* (2018 edition, Publications Office of the European Union 2018).

— *Handbook on European non-discrimination law* (2018 edition, Publications Office of the European Union 2018).

James G and others, *An introduction to statistical learning: with applications in R* (Second edition, Springer texts in statistics, Springer 2021).

O'Neil C, *Weapons of math destruction: how big data increases inequality and threatens democracy* (Penguin Books 2016).

Contributions to collections

Calders T and Žliobaitė I, 'Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures' in *Studies in Applied Philosophy, Epistemology and Rational Ethics* (Springer, Berlin, Heidelberg 2013) vol 3.

Articles

Calders T and Verwer S, 'Three naive Bayes approaches for discrimination-free classification' (2010) 21(2) Data Mining and Knowledge Discovery 277.

Dwork C and others, 'Fairness Through Awareness' [2011] (arXiv:1104.3913 [cs] type: article).

Kamiran F and Calders T, 'Data preprocessing techniques for classification without discrimination' (2012) 33(1) Knowledge and Information Systems 1.

Kamiran F, Žliobaitė I, and Calders T, 'Quantifying explainable discrimination and removing illegal discrimination in automated decision making' (2012) 35(3) Knowledge and Information Systems 2012 35:3 613.

— 'Quantifying explainable discrimination and removing illegal discrimination in automated decision making' (2013) 35(3) Knowledge and Information Systems 613.

Kusner M and others, 'Counterfactual Fairness' [2018].

Mehrabi N and others, 'A Survey on Bias and Fairness in Machine Learning' [2022].

Prince A and Schwarcz D, 'Proxy Discrimination in the Age of Artificial Intelligence and Big Data' en 105 IOWA LAW REVIEW 62.

Stoyanovich J and others, 'Responsible data management' en (2022) 65(6) Communications of the ACM 64.

Wachter S, Mittelstadt B, and Russell C, 'Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI' [2020] (ID 3547922).

Wiśniewski J and Biecek P, 'fairmodels: A Flexible Tool For Bias Detection, Visualization, And Mitigation' [2022].

Zafar M and others, 'Fairness Constraints: Mechanisms for Fair Classification' [2017].

Žliobaitė I and Custers B, 'Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models' (2016) 24(2) Artificial Intelligence and Law 2016 24:2 183.

Other works

30% of European businesses are still not compliant with GDPR, 'RSM Global' (July 2019) <https://www.rsm.global/news/30-european-businesses-are-still-not-compliant-gdpr> accessed 16 July 2022.

Article 29 Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679* (2017).

Autoriteit Persoonsgegevens, *De verwerking van de nationaliteit van aanvragers van kinderopvangtoeslag* (2020).

Balayn A and Gürses S, *Beyond Debiasing: Regulating AI and its inequalities* (2021).

Carlisle M, racist data destruction?, 'Medium' (January 2020) <https://medium.com/@docintangible/racist-data-destruction-113e3eff54a8> accessed 7 July 2022.

Commission, *Proposal for a Regulation laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)* (COM/2021/206 final).

European Union Agency for Fundamental Rights, *'#BigData: Discrimination in data-supported decision making* (2018).

European Union Agency for Fundamental Rights, *Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights* (2019).

Galhotra S, Brun Y, and Meliou A, 'Fairness testing: testing software for discrimination' (ACM August 2017).

Kamiran F and Calders T, 'Classifying without discriminating' (2009).

Python Software Foundation, The Python Language Reference (Version 3.10.5,) <https://docs.python.org/3/reference/> accessed 7 July 2022.

R Core Team, R: A language and environment for statistical computing (2022) <https://www.R-project.org/>.

Verma S and Rubin J, 'Fairness definitions explained' (FairWare '18, Association for Computing Machinery May 2018).

Why do people resist EDI initiatives?, 'Association of Medical Research Charities' <https://www.amrc.org.uk/blog/why-do-people-resist-edi-initiatives> accessed 17 July 2022.