University of Antwerp

Faculty of Science

Targeted Approaches Against Discrimination

New Methods for Bias Detection and Mitigation in Automated Decision Making Systems

Thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy in Computer Science
at the University of Antwerp

Daphne Lenders

Supervisors
prof. dr. T. Calders
prof. dr. S. de Raedt

Antwerpen, 2024

Jury

Chairman
prof. dr. D. Martens, University of Antwerp, Belgium

Supervisors
prof. dr. T. Calders, University of Antwerp, Belgium
prof. dr. S. de Raedt, University of Antwerp, Belgium

Members
prof. dr. B. Goethals, University of Antwerp, Belgium
prof. dr. F. Giannotti, Scuola Normale Superiore, Italy
prof. dr. M. Pechenizkiy, TU Eindhoven, The Netherlands
prof. dr. E. Ntoutsi, Bundeswehr University Munich, Germany

Contact
Daphne Lenders

University of Antwerp
Faculty of Science
Adrem Data Lab
Middelheimlaan 1, 2020 Antwerpen, België
M: daphne.lenders@uantwerpen.be

Dutch title:

Gerichte Aanpak van Discriminatie

Innovatieve Methoden voor Het Opsporen en Verminderen van Vooroordelen in Automatische Besluitvorming

Acknowledgements

It is hard to express the gratitude that I have felt these last four years, both for the opportunities given to me and the people who helped me make the most of them. While I am not going to deny that a PhD can be tough and lonely on some days, on most days it felt like being allowed the time and space to study topics I truly care about, while meeting and learning from an incredible and inspirational community.

I am very grateful to my supervisor, Toon Calders, for providing this time and space and giving me freedom to explore my research interests, while always being available for guidance and feedback. Thank you for your ideas, your patience, your approachability and for always encouraging new research directions and opportunities. Thank you also to Sylvie de Raedt for any legal insights into this work and thank you to my PhD jury: David Martens, Bart Goethals, Eirini Ntoutsi, Mykola Pechenizkiy and Fosca Giannotti. Thank you for your time and efforts in reading this thesis and your interesting feedback and questions during my defense.

Next, I want to thank all the people that I have met on the way of this PhD, during conference visits and summer schools. It is too many names to mention, but I hope they know who they are. Without some shared discussions and critical feedback I would not have developed the ideas that I have now and that made this thesis possible. More important than any academic discussion, I want to express my gratitude towards all the fun moments we shared: the sense of community within the algorithmic fairness researchers has always felt strong and it has been special to see the same familiar faces in different corners of the world: from sharing fries in Leiden, to singing Karaoke in Seoul or watching the Euro Cup in Germany. I will miss those moments and will always look for opportunities to visit another EWAF, another FAccT or another ECML PKDD just to experience them again.

I would also like to thank everyone who has made my time at the Scuola Normale Superiore in Pisa possible and special. Thank you Fosca Giannotti, for welcoming me to the group and thank you, as well as Andrea Pugnana, Roberto Pellungrini and Dino Pedreschi for your efforts in our shared paper. Many thanks also go to Margherita Lalli for sharing her espressos and salads with me and making the lunch breaks memorable. Last but not least, thank you to Océane Duvert, for all the adventures in Italy and all the magical escape moments from work filled with sunsets and aperitivos.

A lot of gratitude also goes to my friends outside of the PhD. Thank you Nynke Zwart for our endless voice message exchanges and for giving friendship that survives both space and time. Thank you Anwar Hiralal for encouraging me to climb both physical and metaphorical fences. Thank you to all the iterations of dear housemates from Harmoniestraat 150: Izaskun, Laura, Abraham, Valerie, Alejandro, Jana, Alessandra and Ivan. Most of my PhD time I have spent living in our shared home, hence any memories

about the last four years will always be connected to memories of green kitchen walls, heated Obama vs. Ronaldo debates, pancake mornings and tortilla nights. Coming home to you and having your companionship made it easy to let go of any work-related stress and just enjoy this period of life.

Next to my friends I want to thank my family. *Mama en papa, dankjulliewel om er altijd voor me te zijn, voor alle Nettersheim bezoeken om te kunnen ontspannen en alle telefoongesprekken, waarin ik me kon afreageren als dat nodig was. Elene, dankjewel voor ons "Skype Office" en alle grappen die daarin gemaakt werden. Jeanne, dankjewel voor alle politieke + feministische discussies die me blijven aanmoedigen om in dit onderzoeksgebied te werken.* Elie, thank you so much for designing this thesis' cover and making it much prettier than I could have.

My final thanks, go to all colleagues within the University of Antwerp, both from the DigiTax and Adrem Data Lab. Again, too many names to mention, but I feel grateful for all shared moments, lunches and discussions and hope that there will be many upcoming Adrem BBQs, where also alumni will be invited! Of all colleagues, special thanks go to Anne Oolo, for our shared paper and to Joey de Pauw for being my office mate and patiently enduring my Daphne-questions. My biggest thanks go to Marco Favier and Ewoenam Kwaku Tokpo: thank you for having had our first paper rejected at the same time and managing to laugh about it, thank you for all the coffee breaks and distractions, thank you for being slightly annoying with your endless debates about capitalism, thank you for being there when I needed it. I am sad to not have you as colleagues anymore but am incredibly grateful to keep you as friends.

Summary

Automated decision-making (ADM) systems used in high-stakes areas such as lending or hiring often perpetuate biases present in their underlying data. Consequently, these systems can adversely impact certain population groups, mirroring the sexist or racist practices of our society. In this thesis, we inspect current approaches to auditing and mitigating such discriminatory biases in ADM systems. We highlight how these approaches typically centre around single definitions of fairness, that aim to express how (un)fair some system is through a single number and try to optimize for fairness accordingly. We explain how these approaches fall short in adequately understanding and resolving discrimination and argue how better approaches should be driven by more nuanced considerations: rather than having one single fairness measure, auditors should focus on which parts of the data a system behaves discriminatory, so that they then can then address this behaviour in a targeted manner. To that end, our first two chapters focus on new tools and methods for bias detection in ADM systems. The first inspects the potential of interactive auditing toolkits, while the second improves an existing method for measuring individual fairness, allowing auditors to decide for one decision subject at a time whether they received just treatment. Our third chapter introduces a human-in-the-loop approach to mitigate bias in ADM systems. We design a selective classifier that refrains from making predictions when they are deemed as discriminatory. These rejected instances, along with an explanation for their rejection, can be passed on to human experts who can make better-informed decisions for them. The fourth chapter shifts focus from new bias mitigation techniques to evaluating their effectiveness. We emphasize how the traditional evaluation scheme, based on single fairness definitions, is not sufficient and instead introduce a benchmarking-dataset to facilitate the evaluation of bias mitigation strategies. This dataset includes a fair and biased version of its decision labels, allowing precise assessment of how well a model can predict the fair labels after being applied on the biased ones. Our fifth and final chapter zooms out from these specific considerations surrounding bias in ADM systems and provides an overview of the research field in general and how it has developed over the last 15 years. By highlighting research gaps, we also conclude this thesis with a discussion and its implications for future work.

Samenvatting

Systemen voor geautomatiseerde besluitvorming (ABV) die worden gebruik in risicovolle domeinen zoals leningen of aanwerving, nemen vaak de vooroordelen van hun onderliggende data over. Hierdoor kunnen deze systemen bepaalde bevolkingsgroepen negatief beinvloeden, en seksistische of racistische praktijken van de samenleving weerspiegelen. In dit proefschrift onderzoeken we de huidige manieren voor het controleren en beperken van discriminatie in ABV systemen. Conventionele methodes, richten zich op enkelvoudige definities van eerlijkheid, die door middel van een enkel getal proberen uit te drukken hoe (on)eerlijk een system is en eerlijkheid dusdaning optimaliseren. We leggen uit waarom dit soort methodes niet geschikt zijn om discriminatie volledig te begrijpen en te bestrijden en hoezo meer genuanceerde overwegingen noodzakelijk zijn: in plaats van één enkele maatstaf voor eerlijkheid te hanteren, moeten we begrijpen op welke specifieke delen van de data een ABV system discrimineert, zodat we dit gedrag gericht kunnen aanpakken. De eerste twee hoofstukken van dit proefschrift richten zich daarom op nieuwe systemen en methoden voor het detecteren van discriminatie in ABV systemen. Het eerste hoofdstuk onderzoekt het potentieel van interactieve auditinstrumenten terwijl het tweede hoofdstuk een bestaande methode voor het meten van individuele eerlijkheid verbetert. Hiermee kan men voor elk beslissingsonderwerp afzonderlijk bepalen of een ABV system een rechtvaardig besluit maakt. Ons derde hoofdstuk introduceert een human-in-the-loop manier om discriminatie in ABV-systemen tegen te gaan. We ontwerpen een selectief besluitvorming model, dat afziet van voorspellingen wanneer deze als discriminerend worden beschouwd. De afgewezen gevallen kunnen, samen met een verklaring voor hun afwijzing, worden doorgegeven aan menselijke experts die betere beslissingen kunnen nemen. In het vierde hoofdstuk onderzoeken we hoe de effectiviteit van discriminatie-beperking technieken geevalueerd kan worden. We laten zien hoe het traditionele evaluatieschema, gebaseerd op enkelvoudige definiets van eerlijkheid, niet voldoende is en introduceren in plaats daarvan een benchmark-dataset om de evaluatie van discriminatie-beperking technieken te vergemakkelijken. Deze dataset bevat zowel eerlijke als bevooroordeelde versies van de beslissingslabels, zodat nauwkeurig kan worden beoordeeld hoe goed een model de eerlijke labels kan voorspellen na toepassing op de bevooroordeelde labels. Ons vijfde en laatste hoofdstuk gaat niet om specifieke overwegingen rondom discriminatie door ABV-systemen, maar geeft een algemeen overzicht van het onderzoeksveld en hoe het zich de afgelopen 15 jaar heeft ontwikkeld. We benadrukken waar huidig onderzoek te kort schiet, en geven op grond daarvan suggesties voor nieuwe onderzoekslijnen die uit dit proefschrift naar voren komen.

Contents

Introduction

Over the past years, an increasing amount of automated decision-making (ADM) systems have been used in high-stake decision areas, such as lending, hiring or recidivism prediction. These systems learn from historical data, coming in the form of tables, text, images other formats, to predict a class label from the data's features. For instance, in a lending setting banks commonly have information about past loan applicants, such as their age, occupation, their credit history etc. Given some recorded label of interest, in this case, whether the applicants were granted loans or not, an algorithm can learn from the data to differentiate between those loan applicants who should be approved and those who should be denied. As a result of this learning process, the algorithm outputs an ADM model that can make predictions for new applicants.

Models like these have the potential to make decision processes more efficient and potentially more accurate. After all, they base their decisions on subtle statistical patterns in the data, picking up correlations between the applicants' characteristics and their chances to receive a loan, that might not be obvious to human-decision makers. Yet, it is precisely this mirroring of statistical patterns that gives rise to one of the biggest risks associated with ADM systems: their perpetuation of discriminatory biases present in the training data [93].

Consider the data that a loan approval system is based on. Even if no errors are made in the recording of this data, it still contains all the systemic biases present in our society. These range from representation bias, wherein only population groups with access to financial services will be recorded in the data, to historical inequalities that link some demographic groups to higher levels of education, income, or other relevant factors in loan allocation settings. Lastly, given the racist, sexist and otherwise discriminatory stereotypes in our society, the data is likely to capture directly discriminatory biases: some people are denied a loan, simply because of their skin colour, their gender, or other demographic factors. It is no surprise that any learning algorithm made to capture the statistical associations between the data and the decision of interest, will pick out these correlations and mirror them accordingly. That this is not just a theoretical threat but a harsh reality harming many individuals' lives has been brought to light by many case studies of discriminatory algorithms. One notorious example is the *COMPAS* case, where a model trained to make recidivism predictions, unjustifiably predicted higher risk scores for black than white defendants [83]. Other examples include the *childcare benefit scandal* in the Netherlands, where false allegations of fraud were disproportionately targeted towards people with immigration backgrounds or dual citizenship [59].

Despite the public attention that these case studies have received, ADM systems are still being increasingly deployed for decision processes in both the private and public sector. It is therefore clear that the detection and ultimately mitigation of their biases is an urgent matter. It is much less clear, however, how to actually approach the problem. In this thesis, we will see how many of the traditional approaches, that measure fairness through a single mathematical definition and mitigate biases in completely automatated ways, fall short of addressing the deeper issues at hand [21, 129]. As an alternative, we will discuss how more flexible and context-dependent approaches, guided by domain experts and ethicists, can allow for a more meaningful understanding and resolution of biases. While we do not want to downplay the importance of stakeholder-driven approaches to this topic, that focus on the involvement of policy-makers, industry representatives and affected communities [16,18,140], we will still discuss this topic from a technological point of view. In other words, we assess how technological systems can be designed, to allow for a flexible and context-driven engagement with algorithmic biases, that empower human auditors to better understand where biases occur and how to fix them. For the largest part of this thesis, we are going to focus on algorithmic fairness in ADM systems based on tabular data (the terms ADM system and classifier will be used interchangeably). For some reflection on the risk of discrimination in other algorithmic systems, we refer to Chapter 6 of this thesis.

The remainder of this introductory chapter is structured as follows: first, we will give a short introduction to current methods to detect and measure bias in ADM systems. In highlighting the disadvantages of these approaches, we will introduce the content of Chapter 1 and Chapter 2 of this thesis, as both chapters introduce more nuanced methods for bias detection. We will then continue with an overview of current research on bias mitigation. We will discuss some existing approaches to make ADM systems fairer and also explain how researchers typically evaluate the effectiveness of these fairness interventions. Related to both of these topics, we introduce Chapter 3 and Chapter 4 of this thesis, as Chapter 3 discusses a less automated approach towards bias mitigation, and Chapter 4 describes a dataset that we have gathered, meant to facilitate the evaluation of fairness intervention strategies. Finally, we introduce Chapter 5 of this thesis, which does not deal with specific bias detection or mitigation methods but provides a high level overview of other research areas and gaps within the field of algorithmic fairness.

Measuring Bias in ADM Systems

To formally assess whether an ADM system is biased towards some population groups, one needs some metric defining what it would mean for the system to be absent of discriminatory biases, i.e. what it would mean for the system to be fair. Over the years multiple fairness metrics have been proposed, that can be roughly divided into two categories: group fairness metrics and individual fairness metrics. Before we are going to describe them in more detail, we are going to introduce some mathematical notation that we will use throughout this section:

- X denotes the entire set of individuals we are making decisions for. Consequentially we will use $\mathbf{x}$ to refer to one individual, for which $\mathbf{x} \in X$.

- A is a sensitive variable (e.g. race, gender, religion, ...) which for now we assume to be binary. We use $A(\mathbf{x}) = +$ to denote that $\mathbf{x}$ belongs to the demographic group we consider to be privileged by society (e.g., men), and $A(\mathbf{x}) = -$ to denote that $\mathbf{x}$ belongs to a non-priviliged group (e.g. women).

- Each individual is associated with a decision outcome D, in case of our running example, it is whether they were granted a loan or not. We use $D = 1$ and $D = 0$ respectively, to denote the favourable and non-favourable decision outcome. For a loan allocation system, e.g. $D = 1$ means being granted a loan.
- We use f for the ADM system supposed to mimic the recorded decision outcome Y. $f(\mathbf{x})$ denotes the decision that was taken by f for $\mathbf{x}$. Again it may or may not be in favour of $\mathbf{x}$ (respectively, $f(\mathbf{x}) = 1$ and $f(\mathbf{x}) = 0$).

Further, we are going to illustrate the different metrics based on a toy example shown in Table 1. This Table displays the decisions of two loan allocation models, that were trained on historical bank data with sexist biases in its original labels.

Table 1: Toy dataset of a loan allocation setting, with two ADM models trained to predict the original decision labels.

#	Sex	Credit Amount	Credit History	....	Bank Decision	Model A	Model B
1	Female	10k	No defaults	....	No Loan	No Loan	Loan
2	Female	40k	Had defaults	....	No Loan	Loan	Loan
3	Male	30k	Unknown	....	No Loan	No Loan	Loan
4	Male	12k	No defaults	....	Loan	Loan	Loan

Group Based Fairness

Group based fairness metrics assess the fairness of ADM systems by comparing their predicted labels over the different demographic groups as defined by A. This comparison can either be made solely based on the decisions of a model f on those groups, or on the errors f makes on them. The first category of group metrics are called *outcome-based metrics*, while the second one are called *error-based metrics*.

Outcome-Based Metrics The most well-known outcome-based metric of fairness is called *Demographic Parity*, which in a setting with a binary sensitive attribute requires there to be no difference in positive decision ratio labels between the privileged group and the underprivileged one (see 1).

$$P(f(\mathbf{x}) = 1 | A(\mathbf{x}) = +) = P(f(\mathbf{x}) = 1 | A(\mathbf{x}) = -) \tag{1}$$

Inspecting our toy problem in Table 1 we see that both ADM models satisfy demographic parity in terms of the sex of loan applicants, as in both models the same ratio of male and female applicants are granted a loan. Additionally, we see that the original decision labels did not satisfy this metric, as 50% more of the male applicants were granted a loan than the female ones. Striving for demographic parity in a loan allocation setting makes sense when considering how receiving a loan opens up many opportunities, like buying a house or starting a business: in a fair world, one would expect these opportunities to be equally distributed across demographics. However, the problem with striving for this metric is, that it assumes that all demographic groups are equally eligible for a loan, despite possible differences that might justify handing out more loans to some groups over others. A more strict version of this metric is therefore provided by *Conditional Demographic Parity*. This metric requires the ratio of positive decisions to be equal across demographics, conditioned on some subset of non-sensitive features $L \in X$ that are imperative to the decision task.

$$P(f(\mathbf{x}) = 1|L = l, A(\mathbf{x}) = +) = P(f(\mathbf{x}) = 1|L = l, A(\mathbf{x}) = -)\forall l \in L \tag{2}$$

Essentially, this metric requires the positive decision ratio across demographic groups only to be equal, if these groups share the same non-sensitive characteristics as described by L. For instance, in our loan application setting, it seems reasonable to impose that loan approval ratios should be equal across all demographics, conditioned on not having a history of defaulting. In this case, we see in Table 1 that Model A does not satisfy demographic parity: as the ratio of approved male applicants without default history is 100%, while it is 0% for female applicants without defaulting history. Model B on the other hand does satisfy conditional demographic parity (100% approval ratio across sexes).

Error Based Measures In some settings, where we want to put more emphasis on the ground truth labels that are available in the data, it makes more sense to inspect error-based fairness metrics. Consider, e.g. the type of errors of a lending decision system. First, there are False Positive errors, describing the ratio of instances that are (according to the ground truth labels) not eligible for a loan, but were decided to be granted a loan by the ADM system anyway. In a fair world, one would expect the False Positive Rates to be similar across demographic groups. This notion is captured by the fairness metric of *Equal Opportunity*, as defined as follows:

$$P(f(\mathbf{x}) = 1|D(\mathbf{x}) = 0, A(\mathbf{x}) = +) = P(f(\mathbf{x}) = 1|D(\mathbf{x}) = 0, A(\mathbf{x}) = -) \tag{3}$$

To exemplify in our toy problem, it is clear that Model A does not satisfy Equal Opportunity: here the False Positive Rate for men is 0% while it is 50% for women. While, overall the False Positive Rates are worse for Model B, it does still satisfy the fairness metric as for both men and women these rates are 100%.

Next to inspecting a system based on its False Positive Rates, we can also assess its False Negative rates, i.e. the ratio of instances not granted a loan even though they were considered eligible for one by past decision-makers. Again, we would want these error rates to be equal across demographic groups, which is what the fairness metric of *Equal Risk* requires.

$$P(f(\mathbf{x}) = 0|D(\mathbf{x}) = 1, A(\mathbf{x}) = +) = P(f(\mathbf{x}) = 0|D(\mathbf{x}) = 1, A(\mathbf{x}) = -) \tag{4}$$

Assessing Table 1 we see both models satisfy Equal Risk, as the False Negatives are equally distributed across men and women.

The fairness metric that requires both *Equal Risk* and *Equal Opportunity* to hold is called *Equal Odds*. In other words, this metric demands that both the False Negative and False Positive rates are independent of the sensitive features in the data.

Individual Fairness Metrics

Different from group-based fairness metrics, individual metrics assess for each decision subject at a time whether their prediction label can be considered fair. This type of metric has evolved out of concerns that group-based metrics only provide “shallow" statistical information on ADM systems' biases, without guaranteeing that individual instances are protected from their discriminatory behaviour [42, 52]. Individual fairness metrics operate on the principle of *treating likes alike*, hence they require that individuals who are similar according to some task-dependent similarity metric (commonly defined as an inverse distance metric [89]) also receive similar decision outputs. For instance, in our loan application setting, we could require that any group of individuals who are similar in terms of their credit amount and their credit history should also receive the same prediction label. Though intuitively this fairness notion makes a lot of sense, finding an appropriate way to define “similarity" is not trivial. Consider instance #2 from Table 1 and imagine assessing the fairness of her prediction label based on her most similar instance, which we can select from Table 2. It is difficult to decide whether instance #5 is most comparable to her, due to their similar credit amount, or instance #6, due to also having a history of defaulting. With bigger datasets and more features to consider, the problem of defining similarity becomes even more evident.

Table 2: Illustration of the difficulty of defining *similarity*, when assessing individual fairness. Consider instance #2 from Table 1. Is #5 more similar to her or #6?

#	Sex	Credit Amount	Credit History	....	Bank Decision	Model A	Model B
5	Female	40k	No defaults	....	Loan	Loan	Loan
6	Male	5k	Had defaults	....	No Loan	No Loan	No Loan

Choosing between Metrics and going Beyond Single Notions

Though we have highlighted only a few of many fairness definitions that have been introduced in the literature over the years, it should already have become apparent that each captures some important intuition behind the meaning of fairness, yet, each comes with a set of disadvantages. Hence, choosing which metrics to consider for the definition of a system's fairness, is far from arbitrary and will depend, among others, on these considerations:

- *How reliable is the ground truth* - Error-based fairness metrics define fairness by comparing the ADM systems' prediction inaccuracies across demographics. However, what is considered to be inaccurate is determined by existing ground truth data, which may be faulty/biased. For example, when bankers assess the eligibility of loan applicants, unconscious biases such as racism or sexism, may influence their decisions. Consequentially, some demographic groups will disproportionately be denied a loan, even if they would have successfully repaid one when given the chance. Failing to acknowledge these biases in historical decision-making processes and solely requiring ADM systems to make equal portions of errors as defined by the ground truth, comes at the risk of perpetuating the biases in them

- *How costly are errors?* - While the ground truth of a dataset can rarely be seen as 100% reliable, it still captures some important information that should be accounted for. An automated lending system that pursues demographic parity, and consequentially grants loans to ineligible candidates, could lead to considerable financial losses for a bank: consider, e.g. Model B in Table 1. While it does satisfy many of the existing fairness metrics, it does hand out a lot of loans that might end up not being repayed. Further, the practice of disproportionately handing out more loans to non-eligible candidates of some demographic group over another can be seen as unfair in itself. For instance, we have already seen how Model A in our toy example can be considered fair in terms of *Demographic Parity*, but is unfair according to *Equal Opportunity*.

- *What domain are we dealing with?* - In some domains ADM systems making errors to satisfy some fairness goal might be more problematic than in others. Consider, for instance, a recidivism prediction algorithm, used to decide whether criminal defendants should be released. Releasing a defendant who ends up recommitting crimes would pose a large risk to the rest of society. Hence, pursuing a strict fairness metric, like demographic parity, is potentially more dangerous in this domain than in others. Also, the domain of a decision task determines how to measure both conditional demographic parity and individual fairness. More precisely, for conditional demographic parity, auditors need to define a domain-appropriate set of attributes to condition on. Similarly, they need to find a suitable distance function for the measurement of individual fairness.

- *Should we prioritize group fairness or individual fairness?* - Solely evaluating a system's fairness according to a group-based fairness metric comes at the risk of overlooking cases of *cherry-picking* [42,52]. This term refers to the random distribution of positive decision outcomes to underprivileged groups to achieve some global fairness goal, without paying further attention to which individuals are granted these benefits. To illustrate, consider Model A in 1. It satisfies fairness in terms of *demographic parity*, yet, when inspecting which of the instances is granted a loan we see a questionable pattern: instance #2, a woman asking for a big credit amount and with a history of defaulting, is granted a loan, while instance #1, a woman without defaulting history and lesser credit, is denied one. This is especially problematic, since for candidate #1 there exists a similar male candidate who was granted a loan, while for #2 this is not the case. In other words, while Model A does satisfy group fairness, it does not satisfy individual fairness - a classical case of cherry-picking.

  While the inspection of individual fairness metrics can account for cases of *cherry-picking*, these metrics alone do not guarantee that certain demographic groups are not consistently excluded from financial opportunities like receiving loans [52], something that can only be measured through group-based metrics.

- *Do we want to transform society?* - Sometimes we want to focus on existing biases not being inflated by an ADM system, other times we want to develop a system that challenges the current status quo of society [128]. Consider again our loan decision task: consistently only handing out loans to a narrow slice of society, does not only exclude others from financially investing in their future but also shapes our image of how an eligible loan applicant looks like. Never giving seemingly less appealing applicants the chance to prove the current system wrong will reinforce our existing stereotypes and widen the socioeconomic gaps of our society. Striving for more strict fairness goals can challenge our current assumptions and beliefs.

These discussion points are not meant to give any conclusive answers on which fairness metrics should be used in which settings. While there have been attempts to answer this question [113], we want to highlight how loaded the choice for an appropriate metric is, and how considering just a single metric to define the fairness of an entire system is rarely sufficient to completely understand other factors playing a role. Recently, researchers have therefore argued for a more flexible approach when auditing the biases of ADM systems. Rather than focusing on demographic parity or equal odds as single metrics, we should try to understand where disparities occur, what their causes are, how they look like on an individual level and to which extent they can be justified [52, 129]. To shed sufficient light on all of these considerations, auditors should not measure a system's fairness through one quick calculation, but instead conduct rigorous and extensive audits where not one but multiple bias metrics are inspected and interpreted under the right domain context. We will explore the idea of more thorough bias audits and possible fairness metrics to consider more deeply in Chapter 1 and 2.

The content of Chapter 1 is based on the following paper:

> Lenders, D., & Calders, T. (2023). Users' needs in interactive bias auditing tools: Introducing a requirement checklist and evaluating existing tools. *AI and Ethics*, 1-29.

In this chapter, we will highlight the potential of interactive toolkits for conducting rigorous bias audits. These toolkits allow auditors to visualize and interact with the model's input data and its predictions. Through this interactive approach, auditors can better understand which parts of the input data a model behaves unfairly on, what some individual examples of discriminatory decisions are, but also assess the biases in the input data itself, such as which demographic groups are underrepresented in it and what other features may be correlated to demographic group membership. Based on a literature review of existing interview studies with auditors and industry practitioners, we identify what requirements tools need to fulfil to be usable in practical and realistic settings. Further, we give an overview of currently existing interactive tools and analyse to which extent they fulfil the identified auditing requirements. By shedding light on how different tools satisfy different requirements, we give concrete suggestions on how some of their functionalities can be combined to create better toolkits. This chapter also sets the scene for the following chapters in this thesis, and highlights the intricate considerations going into audits and their context-dependent nature.

In Chapter 2, we will zoom into the assessment of individual fairness as part of those audits. We explore the already existing Situation Testing algorithm [89], as a method to determine which individuals are discriminatorily affected by an ADM system. This algorithm identifies for some instance in question, the most similar set of “neighbours" from both the privileged and underprivileged group. If the proportion of positive decisions is greater in the former group compared to the latter, it suggests that individuals with similar characteristics, differing only in their sensitive attributes, are not treated equally, leading us to infer discrimination. In Chapter 2, we will address the previously raised concern of finding a suitable distance function that can be used for the similarity analysis. We show how arbitrarily defining some function over all features in the data, without considering their relation to the sensitive attributes or their importance to the decision problem, will produce less reliable fairness assessments. As an alternative, we propose learning a task-dependent distance function based on the available data and highlight the advantages of using this function in measuring individual fairness. We propose utilizing our distance function with situation testing as a tool to flag potentially discriminated instances. These instances can subsequently be forwarded to a human auditor for further examination. The content of Chapter 2 is based on the following paper:

> Lenders, D., & Calders, T. (2021, September). Learning a fair distance function for situation testing. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases. ECML PKDD 2021.* (pp. 631-646).

Fairness Interventions in ADM Systems

Now that we have introduced the considerations that go into detecting the fairness of ADM systems, we discuss some approaches to mitigate possible unfairness issues. We will first focus on completely automated bias mitigation approaches, which can be categorized in one of three categories, describing in which stage of ADM model learning the mitigation takes place: *pre-processing*, *in-processing*, or *post-processing*.

Pre-processing Like the name implies, these methods take place before ADM systems are learned, namely during the pre-processing stage of the training data. The basic assumption is that if the data can be preprocessed such that its discriminatory biases are removed, an algorithm will not have any discriminatory biases to mirror and will be fair from the start. One early example of a pre-processing approach can be found in the work of Kamiran & Calders [74]. They propose to "massage" the data and flip some of the negative decision labels of the underprivileged group to positive while doing the opposite for some positive labels of the privileged group. This is done until *demographic parity* is achieved in the training data. Instances for flipping are selected, according to how close their prediction labels are to the decision boundary of an independently trained classifier on this data. The idea is that we maintain the most information encoded in the original labels, by only flipping instances with high probabilities of belonging to the opposite class.

Another example of data pre-processing lies in the idea of representation learning. The aim is to learn a new representation of the data in which only information relevant to the decision task is encoded, and any information solely relating to sensitive features is disregarded. By training a model on this new representation of the data, its decisions should also not rely on any sensitive information [23, 49, 142]. To exemplify, Calmon et al. propose to frame representation learning as an optimization problem. The objective is to learn a new data representation, that is as close as possible to the original one, but subject to both group and individual fairness constraints. The group fairness constraint imposes a similar positive decision ratio across sensitive groups. The individual fairness constraint ensures that similar individuals are mapped to similar labels in the new data representation, regardless of their sensitive features [23].

In-processing Rather than changing the input data fed to some learning algorithm, it is possible to change the algorithm itself to make its learning process more fair. This is the intuition of *in-processing* methods for bias mitigation methods. Many in-processing methods work by formulating the classification task as a constrained optimization problem, aimed at maximizing some performance measure (e.g., accuracy) subject to some fairness constraint (e.g., demographic parity or equal opportunity) [58, 94, 141]. Adversary methods are another type of in-processing method frequently explored in the literature: the main idea is to simultaneously learn a predictor and an adversary for some decision task. The objective of the latter is to predict the sensitive group membership of some instance, solely based on either the prediction or the prediction error of the former. By simultaneously maximizing the predictors' performance in predicting a decision outcome, and minimizing the adversary's performance in predicting sensitive group membership either demographic parity or equal odds can be achieved (depending on if the adversaries' input is the prediction itself or the prediction error) [144].

Post-processing This last family of bias mitigation methods directly operates on the outputs of an ADM system, while leaving the underlying data and learning algorithm untouched. One common way to do so is by adjusting the prediction probability thresholds for different demographic groups, such that the resulting distribution of prediction labels satisfies some fairness metric of choice [32, 94]. Another method is proposed by Hardt et al., who outline how to flip some of the prediction labels of a model, such that finally equal opportunity is achieved [32, 94].

Measuring Effectiveness of Interventions

After applying one of the fairness interventions throughout the learning pipeline, researchers need to evaluate the effectiveness of those interventions. Given some fairness goal that their fairness intervention was optimizing for, they measure how close their model is to achieving this goal while also evaluating its predictive performance. The common assumption is that to compensate for the unfairness in the original labels, an ADM model must sacrifice some accuracy on them to achieve the imposed fairness constraints. The interplay between predictive and fairness-related measures is commonly summarized as the fairness-accuracy trade-off. Accordingly, a fairness intervention is evaluated as successful if it achieves some fairness goal without compromising the predictive performance too much.

Beyond automated solutions

The fairness intervention techniques that have been discussed here and the way in which their effectiveness is evaluated suffer from important shortcomings, all relating to the automated and non-contextual nature of these approaches. As we have pointed out earlier, measuring the fairness of an ADM model through one single metric is usually not sufficient to understand the biases underlying it. Based on the same reasons, we argue that it is difficult to guarantee fair models when solely optimizing for one of these metrics and not accounting for more nuanced considerations. To reiterate some of the concerns we raised earlier: the choice between which fairness metric to optimize for (and according to which metric to evaluate a system) is far from arbitrary. Picking one metric and blindly fixating on achieving the corresponding fairness goal comes at the risk of ignoring important aspects that are brought to light by other fairness metrics. Related to this, the interventions that only optimize for group-based fairness measures, come at the risk of cherry-picking, i.e. distributing positive decision outcomes arbitrarily across underprivileged groups, without much concern for individual fairness [42, 52].

While some approaches seek to counteract this problem and simultaneously achieve group- and individual fairness (e.g. [23]) by phrasing this as a multi-objective optimization problem, these interventions may be hard to interpret. It becomes difficult to track which fairness metrics are prioritized and which groups of the data receive different decision labels than they would have received under a regular ADM system. Connected to this, there is a general concern that automatized solutions for bias mitigation give little insight into their reasoning and the logic behind where and why interventions are applied [15, 99]. This is both problematic towards the designers of ADM systems, who have little power to incorporate their domain knowledge into the decision-making process [99], and towards the decision subjects, who are left in the dark about why they receive certain decision outcomes [15]. These concerns are also echoed by new legislation around ADM systems, most notoriously the EU AI Act [123]. Article 14 of this act states that high-risk decisions, like loan approvals, cannot be made by automated models alone and instead should be overseen and adaptable by human domain experts.

Drawing forth from this new legislation and the concerns outlined above, Chapter 3 of this thesis will introduce a bias mitigation method that is not fully automatized but instead aims for a hybrid human-machine decision-making process. The proposed methodology is an extension of the selective classification framework, that traditionally has been used for accounting for uncertainty in decision processes. Selective classifiers have the option to reject a given ratio of predictions that they are not certain about and pass these on to human decision-makers or other decision-making models [61]. In Chapter 3, we show how this reject option is an excellent way to account for the unfairness of an ADM model. Using inherently interpretable methods to both check for group- and individual discrimination, our selective classification model rejects instances where it deems its predictions as discriminatory. In our chapter we highlight how the explanations behind these fairness-based rejections can empower a human domain expert to make more well-informed decisions on the corresponding instances, overall increasing the fairness and transparency of the decision process. The content of this third chapter is based on the following paper:

> Lenders, D., Pugnana, A., Pellungrini, R., Calders, T., Pedreschi, D. & Giannotti F. (In press). Interpretable and fair mechanisms for abstaining classifiers. In *ECML PKDD 2024: Joint European Conference on Machine Learning and Knowledge Discovery in Databases*

Questioning the Fairness-Accuracy Trade-Off

As we have explained previously, bias intervention techniques are commonly evaluated according to their fairness-accuracy trade-off, i.e. checking if they satisfy some fairness goal while compromising as little accuracy as possible. We have already highlighted how choosing a single fairness metric and evaluating a bias intervention technique accordingly, is dangerous and comes with the risk of ignoring other relevant fairness concerns, while simultaneously undermining human expertise throughout the decision process.

Beyond that, in this section, we also call into question the logical validity of the presumed trade-off between fairness and accuracy in decision-making. Consider again Table 1 and imagine we want to evaluate both Model A and Model B in terms of this trade-off. Both models satisfy the imposed fairness goal of demographic parity, yet Model A has an accuracy of 75%, while Model B only has an accuracy of 25%. Despite Model A providing a better trade-off, we argue that it is not necessarily objectively better than Model B. The problem here is that we only have access to the historical decisions for each loan applicant, rather than the actual outcome we want to model, i.e. which applicants are eligible and would pay back a loan when receiving one. Imagine a hypothetical world, where we do have access to these labels, as displayed in Table 3. With access to these labels, we can confidently say that Model B is not just the most accurate at predicting these labels, but also the fairest, as all loan applicants get the decision outcome they deserve.

Table 3: Example of the ″fair" decision labels in a loan allocation setting, that do not show whether applicants were granted a loan, but instead highlight if they were eligible for one

#	Deserved Decision
1	Loan
2	Loan
3	Loan
4	Loan

This hypothetical example outlines the problems with evaluating fairness interventions according to the fairness-accuracy trade-off on the labels at hand. The very reason why we are applying fairness interventions in our ADM systems is that we assumed the original decision process to be flawed. Because of biases and errors throughout the process, not every applicant received the decision they deserved and we want to correct these errors. Hence, in trying to measure how well an intervention works, it is internally inconsistent to strive for high accuracy on labels that are not believed to be "true" in the first place. This concern has now been raised by several researchers, both from a theoretical perspective [48, 138] and in experimental settings. Regarding the latter, researchers have simulated data with a fair and biased version of the decision labels and evaluated their fairness interventions by applying them to biased train data and measuring their accuracy on the fair one [51, 138, 144]. Though the advantage of such a controlled setting is obvious, there is no denying that simulated approaches cannot capture the complexity and intricacy of realistic data and its biases [47].

In Chapter 4 of this thesis, we therefore address the challenge of evaluating fairness interventions without relying on synthetic data. To accomplish this, we introduce a novel dataset comprising real-life data of students, including details about their leisure activities and study habits. The dataset includes fair decision labels indicating whether students passed a course. We obtained a biased version of these labels through a human experiment where participants estimated performance based on the students' available information. We show how the latter version of the labels is biased against male students, and how the dataset can be used to evaluate fairness interventions without relying on the logically flawed fairness-accuracy trade-off. Further, we test different bias mitigation algorithms on our sampled dataset and show that interventions that appear to work well according to the traditional evaluation scheme, do not necessarily provide good results in this new setting. While we do emphasize that our dataset should be used with extreme caution and results on it should not be overgeneralized, the experiments in Chapter 4 provide further motivation for changing the way we currently reason about biases in ADM systems. Further, by sampling the biased version of our dataset's labels through a human experiment, we emphasize the importance of humans not just in tackling algorithmic biases, but also in understanding where the biases in our data come from. The content of this chapter is based on the following paper:

> Lenders, D., & Calders, T. (2023, March). Real-life performance of fairness interventions-introducing a new benchmarking dataset for fair ML. In *Proceedings of the 38th ACM/SIGAPP symposium on applied computing* (pp. 350-357).

Zooming out

The majority of this thesis focuses on algorithmic biases in ADM systems and the shortcomings of traditional bias detection and mitigation methods. In each chapter, we tackle concrete challenges relating to this and introduce novel methods for detecting and mitigating biases that provide more flexibility and emphasize the role of humans in these settings.

In Chapter 5 of this work, we zoom out from those specific considerations and provide a high-level overview of other research areas surrounding algorithmic fairness and how the field has developed over the past years. For this purpose, this chapter describes a scoping review of the literature in the field. This review aggregates information from more than 1500 papers dealing with algorithmic fairness, highlighting which domains these papers focus on, what technology they address, and if they focus on particular demographic groups that suffer from algorithmic discrimination. Further, we describe *where* current research efforts are coming from, both assessing the expertise of papers' authors (particularly, if they come from a Computer Science or Law background) and their geographical affiliation. In identifying popular research trends across disciplines, we also highlight which areas around algorithmic fairness remain underexplored. Using some case studies we emphasize the urgency of addressing these topics, also highlighting what unique contributions technological and legal researchers can make. The content of Chapter 5 is based on the following paper:

> Lenders, D. & Oloo A. (*Under submission*). 15 Years of Algorithmic Fairness: Scoping Review of Interdisciplinary Developments in the Field.

Drawing forth from the insights of Chapter 5, the final chapter of this thesis will identify future research areas related to Chapter 1 - Chapter 4 of this thesis and conclude the overall work.

Chapter 1

Interactive Auditing Toolkits

This chapter focuses on detecting biases in Automated Decision-Making (ADM) systems using interactive auditing tools. These tools are designed to be user-friendly, requiring no programming knowledge, yet enabling rigorous and thorough bias audits. To account for the subtle and complex way in which discriminatory patterns may unfold, interactive tools need to offer a wide range of functionalities to ensure that auditors can detect, understand, and contextualize all the important biases within a model. Many interviews and usability studies have been conducted to identify the functional requirements an optimal tool should satisfy. Yet, no extensive checklist of these requirements exists, nor is it clear to which extent current auditing tools fulfil them. In this chapter, we provide an overview of currently existing tools, while also encapsulating auditors' functional needs for such tools in one comprehensive checklist. More importantly, we will evaluate each of the existing tools according to this checklist and identify ways their shortcomings can be overcome. Common points of improvement we identified using our checklist, concern the tools' functionality to let users detect complex forms of bias (like intersectional bias) and let users understand the causes of this bias[1]

1.1 Introduction

One essential part of creating fair ADM systems lies in auditing these systems for potential biases before they are deployed. This need is emphasized by researchers [35, 128] and legal institutions. For instance, Local Law 144 in New York City [8] mandates that any automated hiring/employment system needs to be audited and similar regulations on a more general level are proposed in the EU, in the form of the EU AI Act [46].

While it is clear that bias audits *should* take place, much less is clear about *how* they should be conducted, as the legal regulations only give minimal guidelines about this [35, 111]

[1] This chapter is based on: *Lenders, D., & Calders, T. (2023). Users' needs in interactive bias auditing tools introducing a requirement checklist and evaluating existing tools. AI and Ethics, pages 1-29..*

(for a more elaborate discussion of the regulations we refer to section 1.5). Given these legislation gaps, there are still many open questions, around which components of a model should be audited (e.g., one could only inspect the models' output or also evaluate the code behind the model) or when an audit can be considered to pass or fail [35]. While these are important ongoing debates, the focus of this chapter will lie in the most common part of an algorithmic bias audit: the assessment of the fairness/accuracy of a model's prediction and its underlying data [35]. For simplicity's sake, we will (without wanting to diminish the importance of other parts) use the term "audit" to refer to this specific part of the process. Further, we will use the term "auditor" to refer to the organization/person conducting this assessment. While the discussion on who this auditor should be is an important one, we push these concerns aside and merely assume that an auditor is a team or person (that could either be internal or external of the organization whose system is audited) who wants to detect and understand the biases of an ADM model.

Even when making these simplified assumptions conducting an audit is still complex and challenging, as there are many causes for why an ADM system may be biased, many subgroups that may be affected by it, and many ways in which this bias may unfold (see for instance some of the bias definitions in the introduction of this thesis). Hence, any bias audit should be a rigorous and thorough process that sheds light on all these concerns. To facilitate this process, various tools have been developed to assist auditors. These tools enable them to visually and interactively inspect the underlying data behind an ADM model, as well as its prediction outcomes on new data (e.g., granting a loan or not) [3, 22, 116, 133, 135]. Unlike tools that come in the form of programming libraries these interactive tools are usable by a wide range of people, as their usage does not require technical or coding skills. Additionally, these tools can help standardise the auditing process, by providing clear pointers on which considerations need to be made throughout, and on which potential unfairness issues to explore. Lastly, these tools can have a broad impact because they are accessible to the public for free. They can save time and money for users who don't have to start audits from scratch but can use the tools' existing functionality. Despite their clear potential, many tools are developed in isolation of those who might use them, begging the question of how suitable they are in realistic settings. Interview studies with developers and other possible auditors can reveal an answer to this question.

Veale et al., Holstein et al. and Constanza-Chock et al., for instance, conducted interviews with practitioners and auditors to identify their current procedures in testing the fairness of ADM systems [35, 65, 126]. Though they did not directly explore how interactive tools can aid this process they still identified common obstacles that they face, that should be considered when designing auditing toolkits. For instance, they found that auditors often do not have information about decision subjects' sensitive attributes, like gender or race, complicating the asessment of how a system might impact demographic groups differently. Hence, this reveals the requirement that interactive toolkits should enable a bias audit when sensitive/demographic information is not available.

More recently, other interview studies were conducted in which potential auditors were directly asked to list their requirements in auditing toolkits and identify points in which to improve current ones [84, 86, 98, 110]. The studies identified essential user needs, including the requirement for scalable bias audit tools. Since ADM models may make occasional errors, auditors want to avoid wasting time on random mistakes and focus on significant issues that indicate systematic discriminatory patterns [84].

All aforementioned studies uncover important considerations that should go into the design of interactive tools. However, so far only two attempts have been made to give an extensive overview of all these requirements [98,110]. First, Richardson et al. introduced a rubric listing both functional and non-functional criteria for interactive toolkits [110]. However, this rubric is not specifically targeted to ADM systems and some of its items, like "[tool] can detect bias" or "[tool] contextualizes fairness" remain somewhat vague, and do not provide actionable and concrete suggestions on how to implement them. Nakao et al. also introduce a list of tool requirements. However, they base this list solely on the results of their interview studies and therefore miss essential design needs identified by other research works [98].

In this chapter, we present a literature review of interview studies with practitioners to provide a more complete list of tool requirements. Further, we give concrete and actionable insights into how these requirements can be implemented. We do so by first examining how some currently available toolkits already fulfil some design criteria so that developers of new tools can draw inspiration from their functionality. Second, if none of our examined tools satisfies a given requirement, we provide pointers to relevant literature that gives insight into how some functionality can be implemented. In doing so we are, to the best of our knowledge, the first to provide a detailed overview of some of the interactive tools that are already available.

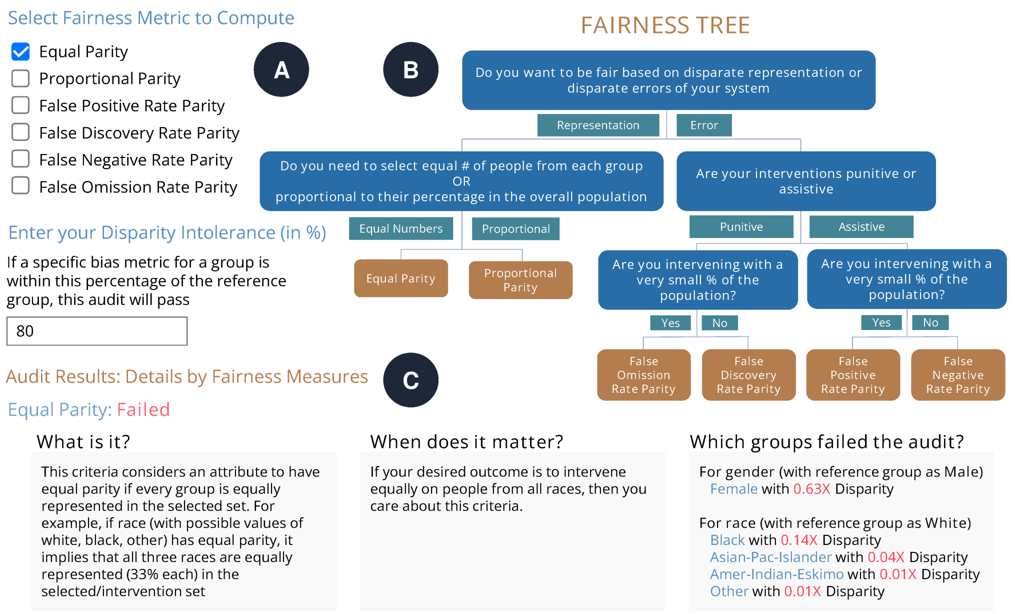
1.2 Overview of selected fairness tools

To give an idea of some of the interactive tools that currently exist, we introduce six tools that we will, later on, evaluate to determine whether they meet the needs of potential auditors. By describing these tools, we do not aim to give a complete overview of all the tools that are currently available but to give an initial understanding of their functionalities. This understanding will form the basis for determining how auditors' functional requirements can be met on a technical level. All the tools we review have an interactive graphical interface, meaning that we excluded tools like AIF360 [12] or FairLearn [95] that come in the form of a Python library. We also only review tools that we were able to use and test ourselves.

To better understand how each tool can be used, we show how each of them assesses bias in the prediction task associated with the "Adult Income Dataset". This dataset contains information on individuals' demographics and working life, like their type of job and their amount of working hours. The associated decision task is to predict whether an individual has a high or low income. We refer to the former as the "favourable outcome" or the "positive label". The dataset contains the attributes "sex", "race" and "age", which are known to elicit biases in ADM models trained on it. We will use the term "protected group" to refer to the group of people that are, based on their sensitive attribute values, historically at lower risk of receiving the favourable outcome than the "unprotected" group. The auditing toolkits described in this section can be used to detect and understand these patterns.

Note that some of the tools are merely prototypes that work only on this *Adult dataset*. Even though they may not be used in real bias audits yet, we discuss their most important components to see how their functionality may be useful to incorporate into future tools.

1.2.1 Aequitas


A
B
C
FAIRNESS TREE
Do you want to be fair based on disparate representation or disparate errors of your system
Representation
Error
Do you need to select equal # of people from each group OR proportional to their percentage in the overall population
Are your interventions punitive or assistive
Equal Numbers
Proportional
Punitive
Assistive
Equal Parity
Proportional Parity
Are you intervening with a very small % of the population?
Are you intervening with a very small % of the population?
Yes
No
Yes
No
False Omission Rate Parity
False Discovery Rate Parity
False Positive Rate Parity
False Negative Rate Parity

Select Fairness Metric to Compute

☑ Equal Parity
☐ Proportional Parity
☐ False Positive Rate Parity
☐ False Discovery Rate Parity
☐ False Negative Rate Parity
☐ False Omission Rate Parity

Enter your Disparity Intolerance (in %)

If a specific bias metric for a group is within this percentage of the reference group, this audit will pass

80

Audit Results: Details by Fairness Measures

Equal Parity: Failed

What is it?

This criteria considers an attribute to have equal parity if every group is equally represented in the selected set. For example, if race (with possible values of white, black, other) has equal parity, it implies that all three races are equally represented (33% each) in the selected/intervention set

When does it matter?

If your desired outcome is to intervene equally on people from all races, then you care about this criteria.

Which groups failed the audit?

For gender (with reference group as Male)
Female with 0.63X Disparity

For race (with reference group as White)
Black with 0.14X Disparity
Asian-Pac-Islander with 0.04X Disparity
Amer-Indian-Eskimo with 0.01X Disparity
Other with 0.01X Disparity

Figure 1.1: Visualization of Aequitas: **(A)** After uploading their data and specifying the sensitive attributes (e.g. "gender" and "race") and the reference groups for these attributes (e.g., "men" and "white"), users can select one or more fairness goals as well as a threshold t. Aequitas will then check for each non-reference group, whether the chosen metric does not diverge more than $(1-t)$ from the reference group. **(B)** The "fairness tree" presents a flowchart that is meant to help users in the choice of the fairness metric **(C)** Extract of the bias report: for every fairness metric selected in (A) it is shown whether the model satisfies this metric or not for the given sensitive attributes. In this case we see that the model does not satisfy "Equal Parity" for race nor gender. Along the attribute "sex" it e.g. shows that there is a disparity of 0.63, meaning that the ratio of men and women predicted to have a high income is 1:0.63. An explanation on how the fairness metric is calculated and why it may matter is also provided.

Aequitas is a web application that can create bias reports, showing for which groups within a dataset an ADM model satisfies some fairness definitions of choice [115][2]. A visualization of Aequitas' interface is given in Figure 1.1.

1.2.2 FairVis

The goal of FairVis is to let users identify intersectional subgroup bias within a model [22]. To this end, users can either compare performance and fairness measures between self-generated subgroups, or explore subgroups on which a model performs unfairly, that

[2] Aequitas also comes in the form of a Python library, with more complex functionalities to detect and mitigate bias in ML algorithms. Since this library can only be used by an experienced programmer, we will only focus on the web application in this paper.

are automatically suggested by the tool.

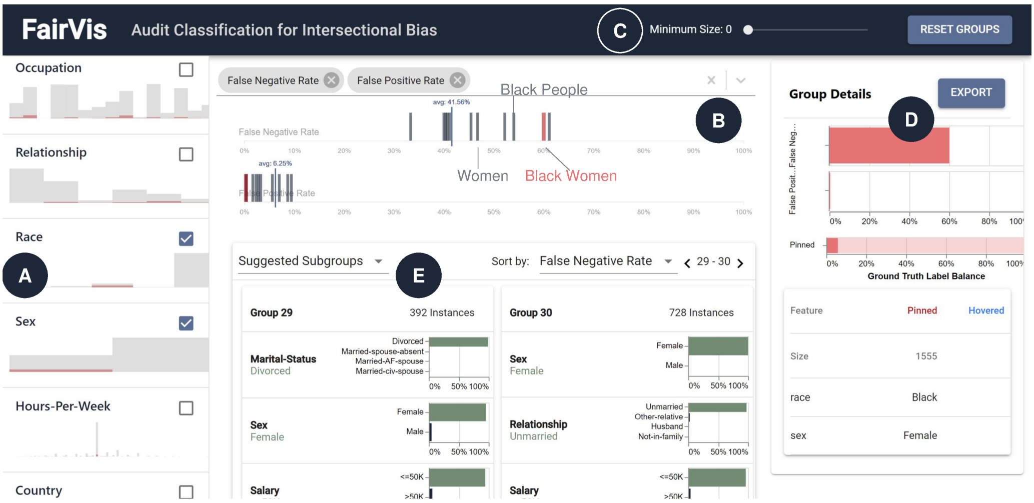


Figure 1.2: Interface of FairVis **(A)** Users can select the attributes they want to generate subgroups for; in this case subgroups based on all possible value combinations of "sex" and "race". **(B)** After selecting some performance metrics (e.g., False Negative Rate), the scores of all subgroups (as generated in (A)) on this metric are visualized. Here we see (among others) the model's False Negative Rate for the population of women, black people and black women specifically, whereas the metrics is highest for black women. Using the slider in **(C)**, users can filter out subgroups smaller than a specified size. **(D)** Here users can get additional information on up to two subgroups as selected in (B), namely the number of individuals belonging to the selected subgroups and the positive decision ratio for them in the data. Here the user sees that there are a total of 1555 black women in the dataset, with a positive decision ratio of ~10% in the labels **(E)** Possible subgroups of interest are suggested to the user, sorted according to their score on a performance/fairness measure of choice. Here groups with a high False Negative Rate are suggested, among others, the groups of divorced and unmarried women.

1.2.3 DiscriLens

The aim of DiscriLens is to visualize discriminatory itemsets, which in this tool are defined as subgroups from the data for which the fairness measure of "conditional demographic parity" is not met. To formally define discriminatory itemsets, assume that we have one sensitive attribute S, one decision variable Y and a set of resolving attributes r. The discriminatory itemsets are then all sets where the conditional demographic parity, defined as $P(Y = 1|S = 1, r) - P(Y = 1|S = 0, r)$, is higher than a threshold τ. In the case of the*Adult dataset*, education and the amount of working hours could be seen as resolving attributes: if an unprotected and protected group (e.g. men vs. women) have the same values on these attributes, but still do not receive a similar ratio of positive decision outcomes, then these group of men and women together constitute a discriminatory itemset.

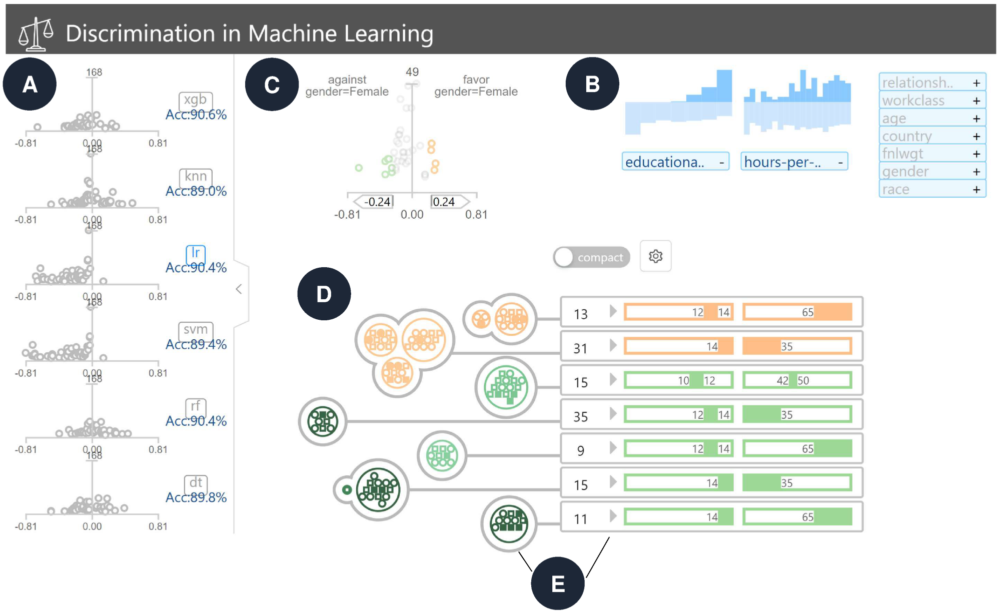
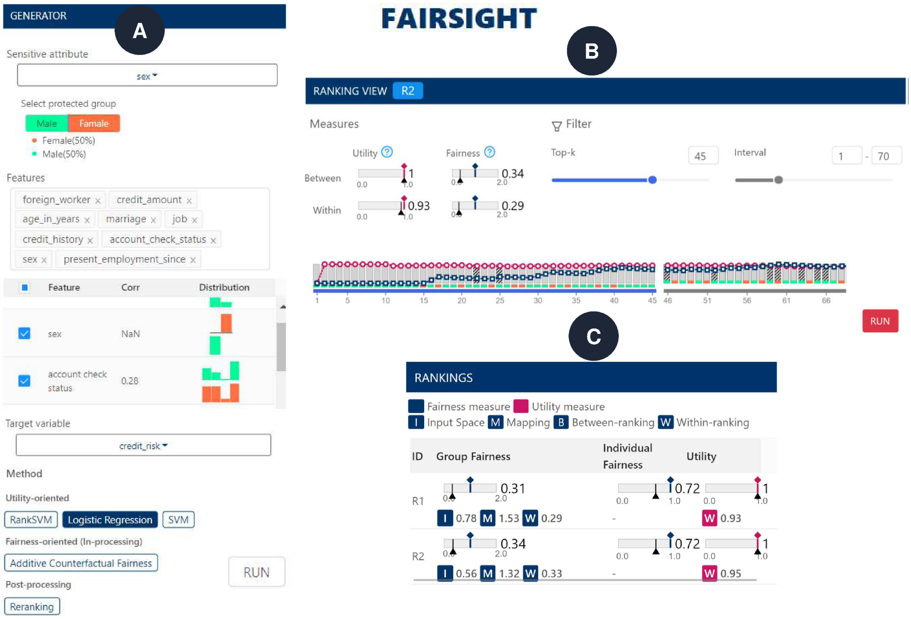


Figure 1.3: Interface of DiscriLens **(A)** Users select the ML model from which they want to inspect the predictions **(B)** Users can select resolving attributes **(C)** Visualization of all discriminatory itemsets, where discriminated itemsets are visualized in green, and favoured ones in orange **(D)** All discriminatory itemsets are here visualized through a so-called 'Ripple Set', which encodes information about the direction of discrimination (either in favour or against a protected group), its severity, and its significance **(E)** One of the discriminatory itemsets; in this case the discriminatory itemset consists of people with a higher education level than 14 and more than 65 workinghours. Within this group there are 5 women (visualized by circles) and 6 men (visualized by squares). The fill colour of the shapes denotes whether the individuals received a positive or negative prediction outcome. In this case, 4 squares are filled, meaning that 4 men received a positive decision label, while none of the women received one (as none of the circles is filled). Because of this high discrepancy, the group is marked as a discriminatory itemset.

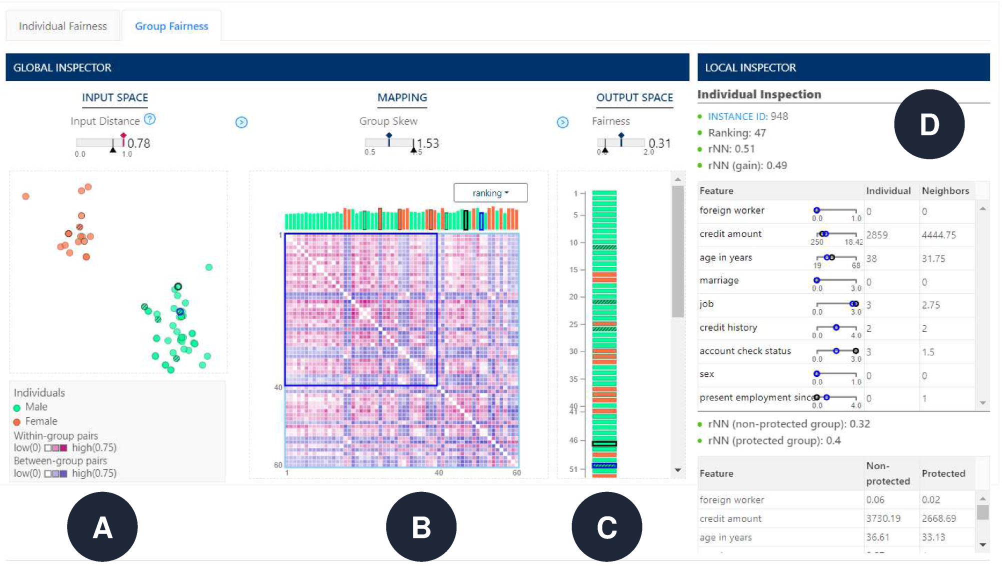
1.2.4 FairSight

The goal of FairSight is to let users identify bias in all three stages of the ML pipeline [3], which are defined as follows: the first stage is the "Input", i.e. the data itself and how it may be differently distributed among the protected and unprotected group. The second stage is the "Mapping", which relates to how the input is mapped to the output and whether similar input data receives similar outcomes. The third stage is the "Output", i.e. the ML model's predictions and how they may be different for protected and unprotected instances. Note that, different from the other tools examined in this paper, FairSight does not operate on binary outcomes of a decision task, but on the rankings generated by a classifier. A classifier's ranking is obtained by ordering the instances according to their assigned prediction probabilities. Assuming that only the top *k* highest ranked individuals get assigned the positive decision label, bias occurs if there are more instances of the unprotected than of the protected group within this top *k*. In the following figures, we give a visualization of FairSight's interface. Note that FairSight is currently a prototype, that only works on the German Credit dataset (and hence it is the only tool not working on our running example of the *Adult dataset*). This data consists of information on loan applicants (incl. sensitive information about, e.g. people's age or gender) and the decision label indicates whether a person was approved for a loan or not. In Figure 1.4, 1.5 and 1.6 we give visualizations of the interface of FairSight.



ID	Group Fairness	Individual Fairness	Utility
R1	0.31 I 0.78 M 1.53 W 0.29	0.72 -	1 W 0.93
R2	0.34 I 0.56 M 1.32 W 0.33	0.72 -	1 W 0.95

Figure 1.4: **(A)** In this "Generator" tab users can select which ADM model to train on the data, as well as the features this model should be trained on. In this case, a Logistic Regression classifier is trained on the features, where "sex" is seen as the sensitive feature (and women are defined as the protected group). Some extra information per feature is given, showing through two histograms how they are distributed differently for the protected and unprotected group **(B)** This is the "Ranking View" tab. After a model has been trained, users can see a visualization of the models' generated ranking, presenting each instance within the ranking as one rectangle, and colouring it according to its protected group membership and its ground truth label (i.e. negative or positive). In this tab, users can also choose a value for k to denote which top-k individuals from this ranking will be assigned a positive label. Some performance and fairness measures are given as well, to show how accurate the ranking is and how fair it is in regard to the proportion of protected and unprotected group members represented in it. **(C)** Here a log is kept of all the ML models the user-generated in **(A)**, and their most important performance and fairness measures are summarized.



Individual Inspection

- INSTANCE ID: 948
- Ranking: 47
- rNN: 0.51
- rNN (gain): 0.49

Feature	Individual	Neighbors
foreign worker	0	0
credit amount	2859	4444.75
age in years	38	31.75
marriage	0	0
job	3	2.75
credit history	2	2
account check status	3	1.5
sex	0	0
present employment since	0	1

- rNN (non-protected group): 0.32
- rNN (protected group): 0.4

Feature	Non-protected	Protected
foreign worker	0.06	0.02
credit amount	3730.19	2668.69
age in years	36.61	33.13

Figure 1.5: **(A)** This "Input Space" inspector is meant to show any fundamental differences between the protected and unprotected group, by visualizing a dimensionality-reduced version of the input data, using different colours for both groups (green for men, red for women). In this case, we see that both groups are quite distinct in their input data, indicating that there may be many features correlated to peoples' sex **(B)** This "Mapping" graph visualizes the so-called "distortion" for each pair of input instances. We speak of high distortion when two instances have similar features on the input space, but received different outcomes. The colours encode the degree of distortion and whether two individuals differ in their sensitive attributes. A dark purple colour between two input pairs, for instance, means that the instances have different sensitive attribute values (i.e. denoted by purple rather than pink colour) and the distortion between them is high (denoted by high colour saturation). A user can inspect this graph to get a high-level overview of the distortion within an ML model, and to select individual instances to inspect more closely (see (d)) **(C)** Similarly, as in Figure 1.4 **(B)** the output ranking of the given ML model is visualized, colour encoding the sensitive attribute of each ranking instance and their ground truth label. We see that a lot more male than female instances are included in the ranking (denoted by colour) and that some of the male instances receiving a positive outcome by the model did not have a positive decision label in the ground truth (denoted by stripes through a block) **(D)** By clicking on one of the instances visualized in (C) users can inspect this instance more closely, and find some measures on how this features ranking position relates to that of its nearest neighbours. If an instance scores low in the ranking, while its most similar neighbours score high, this might be a sign of individual discrimination

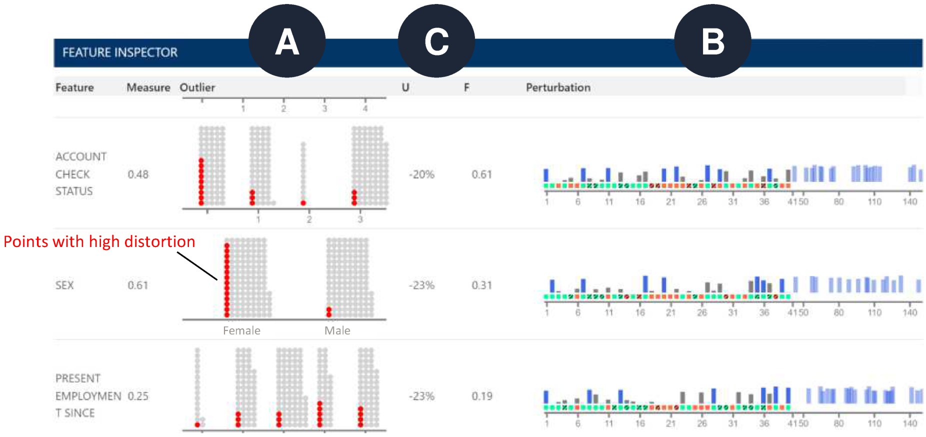
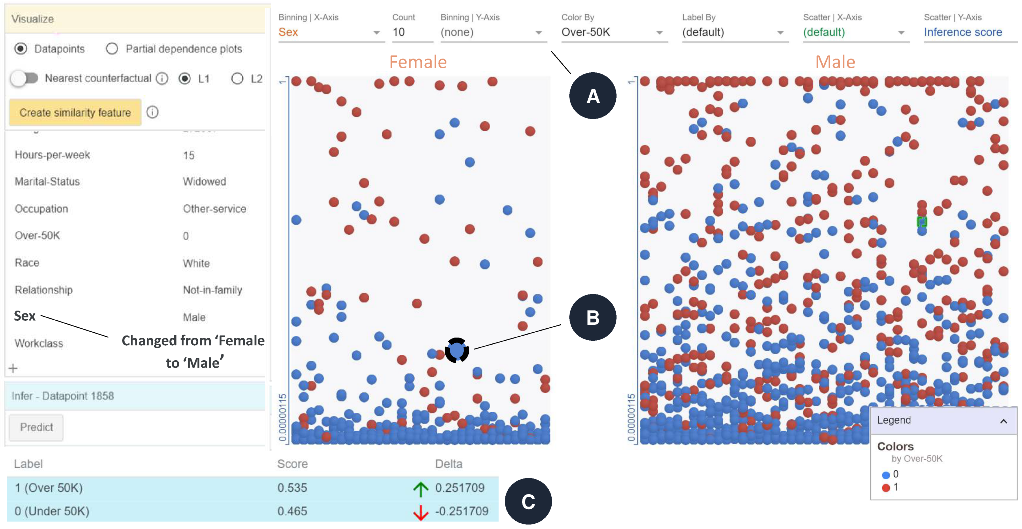


Figure 1.6: Visualization of the "Feature Inspector" tab of FairSight. **(A)** For each feature a histogram is visualized. Within each histogram, data points with the highest distortion in the model (i.e. instances whose nearest neighbours receive significantly different prediction outcomes than the instance itself) are coloured red. Auditors can use this, to inspect whether specific feature values are connected with high distortion for individuals with those feature values. For instance, we see that when splitting people by their "sex", most instances with high distortion, are women, indicating that they are the group suffering most from individual discrimination. **(B)** Here the output ranking is visualized, which is obtained when training a model with a perturbed version of the given feature. Here users can check how perturbing the feature (i.e. removing all correlation between the given attribute and the decision attribute) affects the accuracy of the ranking (i.e. how many individuals who are part of the top-k ranking, also deserve a positive decision label), as well as the fairness of the ranking (i.e. how many protected and unprotected individuals are represented in it). Formal measures of this are provided in **(C)**. Here we for instance see, that when perturbing the attribute "sex" the accuracy of the ranking drops by 23%, while the fairness increases by 0.31

1.2.5 The What-If Tool

The What-If Tool was developed to give users a better understanding of ML models in general, but also specifically in regards to bias in these models [135]. The interface consists of two main components: the "Datapoint Editor" tab (see Figure 1.7) and the "Performance & Fairness" tab (see Figure 1.8). Auditors can use the former to visualize data and model predictions, and select data points to obtain further information or conduct individual fairness analysis on. In the latter, users can inspect a model's performance and fairness on subgroups of choice.



Visualize

● Datapoints ○ Partial dependence plots

Nearest counterfactual ● L1 ○ L2

Create similarity feature

Hours-per-week	15
Marital-Status	Widowed
Occupation	Other-service
Over-50K	0
Race	White
Relationship	Not-in-family
Sex	Male
Workclass	

Changed from 'Female to 'Male'

Infer - Datapoint 1858

Predict

Label	Score	Delta
1 (Over 50K)	0.535	↑ 0.251709
0 (Under 50K)	0.465	↓ -0.251709

Binning | X-Axis
Sex

Count
10

Binning | Y-Axis
(none)

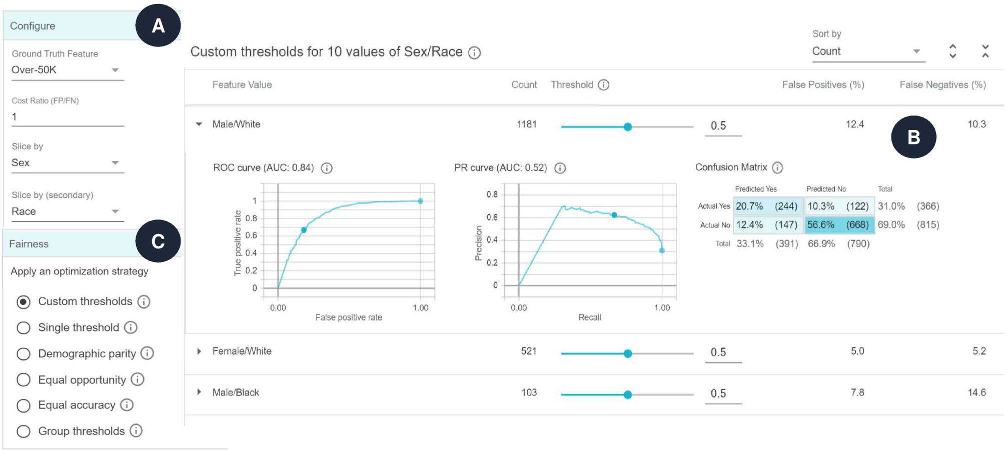
Color By
Over-50K

Label By
(default)

Scatter | X-Axis
(default)

Scatter | Y-Axis
Inference score

Figure 1.7: Interface of the "Datapoint Editor" component of the WhatIf tool. In **(A)** users can select which attributes of the data they want to visualize, using the x- and y-axis in a 2D graph, as well as the colours and labels for each datapoint. In this case "sex" is plotted on the x-axis, and prediction probabilities on the y-axis. This results into two graphs; one for women and one for men. Each point stands for one instance, and their colour denotes their ground truth label (high income - red, low income - blue), while their point in the y-axis denotes their obtained prediction probability (points on the x-axis are randomly spread to increase readability). Here we see, that generally fewer women are represented in the dataset, and it appears that women are more likely to obtain low prediction probabilities. In the graph, users can also select individual data points to either run a counterfactual or a what-if analysis on. (B) In this case, an auditor has selected a female datapoint, to observe how her prediction probability for a high income would be different if her sex was "male" instead. In **(C)** we see that the change in sex increases the prediction probability for a high income by ~ 0.25, indicating a potential case of individual discrimination



Ground Truth Feature: Over-50K

Cost Ratio (FP/FN): 1

Slice by: Sex

Slice by (secondary): Race

Apply an optimization strategy

- ● Custom thresholds
- ○ Single threshold
- ○ Demographic parity
- ○ Equal opportunity
- ○ Equal accuracy
- ○ Group thresholds

Custom thresholds for 10 values of Sex/Race

Feature Value	Count	Threshold	False Positives (%)	False Negatives (%)
Male/White	1181	0.5	12.4	10.3
Female/White	521	0.5	5.0	5.2
Male/Black	103	0.5	7.8	14.6

Confusion Matrix

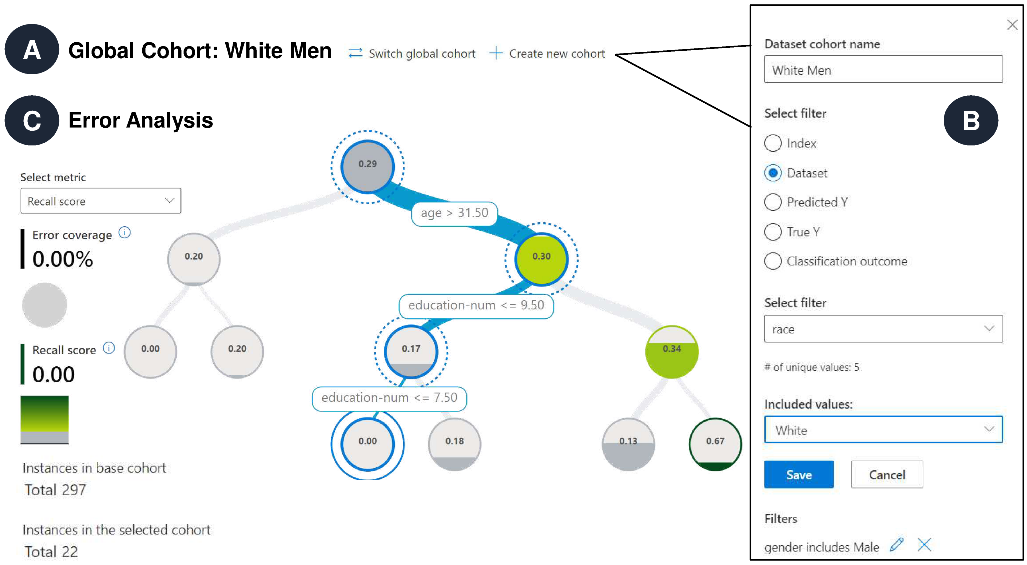
	Predicted Yes		Predicted No		Total	
Actual Yes	20.7%	(244)	10.3%	(122)	31.0%	(366)
Actual No	12.4%	(147)	56.6%	(668)	69.0%	(815)
Total	33.1%	(391)	66.9%	(790)		

Figure 1.8: Interface of the "Performance & Fairness" component of the WhatIf tool. In **(A)** users can select up to two attributes to generate subgroups for (e.g., subgroups based on all value combinations of "sex" and "race"). In **(B)** some performance and fairness measures, as well as the model's ROC curve and confusion matrices for each subgroup are displayed. Users can order these subgroups according to their size or according to one of the fairness/performance metrics. Here, groups were ordered according to size and we can see that the group of white men is most represented in the *Adult dataset*. In this tab, we can also see that they have a higher False Positive rate than all other subgroups. In **(C)** users can try to mitigate the fairness of the model, by adapting the decision threshold to translate a model's prediction probabilities into binary labels. This threshold can either be the same for all subgroups, or differ between them to optimize for a fairness goal of choice.

1.2.6 The Responsible AI Dashboard

The goal of the Responsible AI (RAI) dashboard is to let users understand a model's errors and behaviour, either for the dataset as a whole or for specific subgroups of the data (which can be generated by the user) [116][3]. The dashboard consists of five main features, which we will each discuss separately: in Figure 1.9 we show the "Error Analysis" functionality of the tool, in Figure 1.10 the "Model Statistics" and "Data Explorer" tabs, and in Figure 1.11 the "Feature Importances" and "What-If Counterfactuals" components.

[3] The RAI Dashboard is part of a larger "Responsible AI Toolbox", which also contains functionalities for bias mitigation. Since, however, these functionalities can only be used by experienced programmers we will only focus on the RAI dashboard (that can also be used by non-technical users) in this paper.



A Global Cohort: White Men

C Error Analysis

Select metric: Recall score

Error coverage 0.00%

Recall score 0.00

Instances in base cohort Total 297

Instances in the selected cohort Total 22

Dataset cohort name: White Men

Select filter: ○ Index ● Dataset ○ Predicted Y ○ True Y ○ Classification outcome

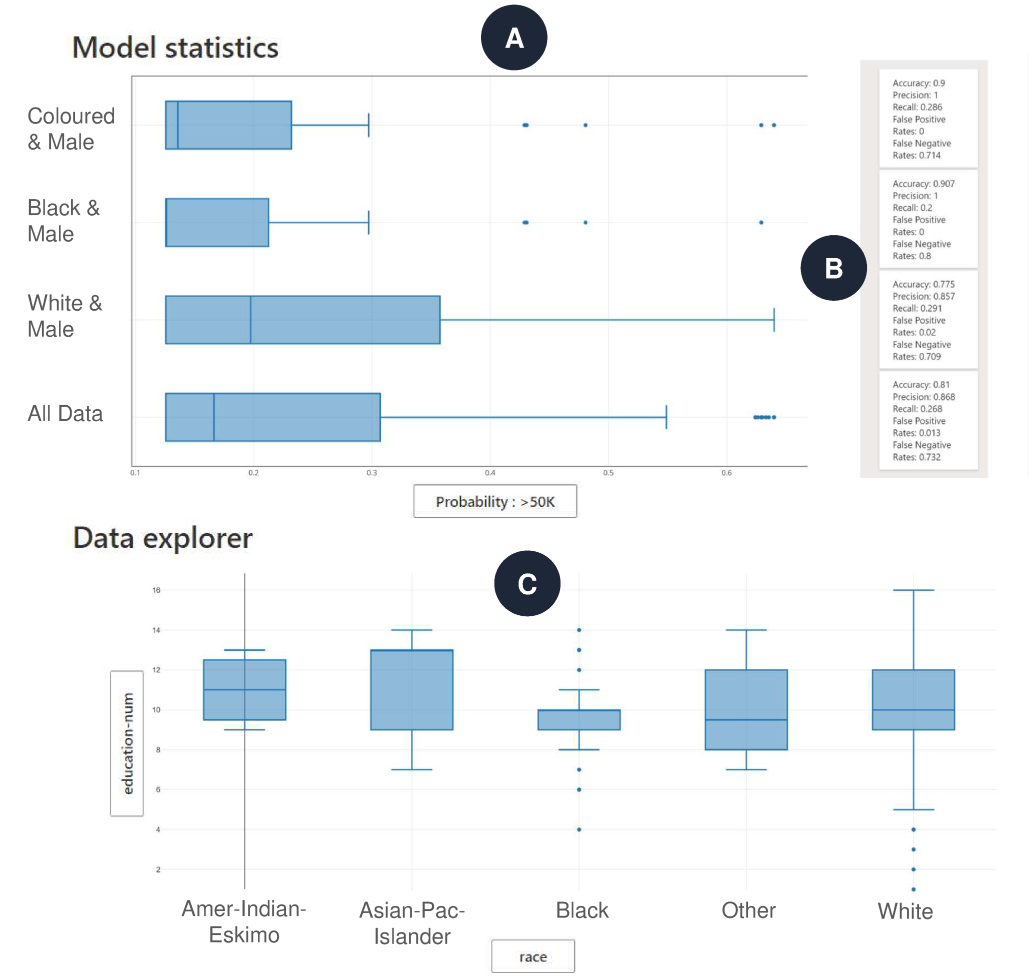
Select filter: race

# of unique values: 5

Included values: White

Filters: gender includes Male

Figure 1.9: Visualization of the error analysis component of the RAI dashboard **(A)** The heading shows which part of the data is currently inspected, either the "global cohort" (i.e. all of the data), or specific subgroups from the data, that a user can generate using the interface shown in **(B)**. In this case the user is inspecting the group of "White Men" from the data. As we see, this group consists of a total of 297 instances. **(C)** Visualization of the "error tree", showing how the errors of the ML models are distributed over different subgroups. Given the subgroup of the data the user wants to inspect and a performance measure of choice, it is visualized on which partitions of the subgroup the performance measure is particularly high or particularly low. In this case, we observe the "recall" performance measure, which is over the whole subgroup of white men 0.29. Following one of the paths of the tree, we see that it is with 0.00 even lower for the group of white men older than 31, with a higher education level smaller than 7 (whereas an education level of 7 indicates the completion of 11th grade in High School). In a similar fashion, a user can inspect subgroups with even higher or lower recall measures.



Accuracy: 0.9
Precision: 1
Recall: 0.286
False Positive Rates: 0
False Negative Rates: 0.714

Accuracy: 0.907
Precision: 1
Recall: 0.2
False Positive Rates: 0
False Negative Rates: 0.8

Accuracy: 0.775
Precision: 0.857
Recall: 0.291
False Positive Rates: 0.02
False Negative Rates: 0.709

Accuracy: 0.81
Precision: 0.868
Recall: 0.268
False Positive Rates: 0.013
False Negative Rates: 0.732

Figure 1.10: **(A)** The "Model Statistics" component of the RAI dashboard, where users can explore the performance of the model on the different subgroups. They can either visualize the distribution of prediction outcomes, of prediction probabilities or of the ground truth labels. Here we e.g. see how the prediction probabilities are distributed for men of different races. The boxplot clearly shows that for white men prediction probabilities are higher than for, e.g., black men. Additionally, in **(B)** some performance measures per subgroup are shown, reflecting patterns we saw in (A). **(C)** The "Data Explorer" component: for a subgroup of choice, users can use the x- and y-axis of the graph to visualize the relation between two features for that subgroup. Here we, e.g. see, how education levels differ among races and how people from asian-pacific-islander background, have for instance higher education levels than black people. Studying these patterns, auditors can reason about inequalities present in the input data.

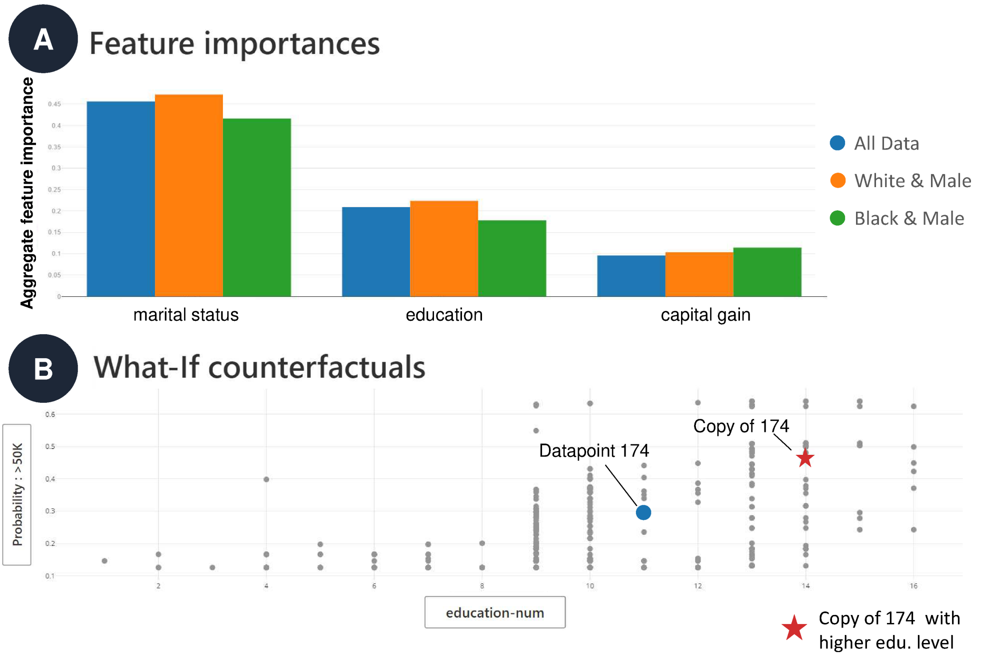


Figure 1.11: **(A)** The top k globally most important features for the decision problem are visualized per user-generated subgroup. In this case, we see that for all subgroups (i.e. the complete data, the group of white men and the group of black men) the same three features are most important for being predicted a high income **(B)** This is the "What-if" analysis tab. Here users can visualize the data according to two attributes of choice and select individual instances to run a what-if analysis on. With this, they can test how changing one or multiple feature values of an instance affects the model's prediction probability for it. Here we can see, for instance, how changing the education level of a data instance positively affects its probability for a positive decision outcome

1.3 Design Considerations for Fairness Tools

While we've only covered a few fairness audit tools, it is evident that interactive tools are developed with different use cases and user needs in mind. However, research has shown that there is a gap between what auditors need from a tool and what functionality the tools offer [86, 110]. To identify this gap, we conducted a literature review to gain insight into auditors' practices and needs that should be accounted for in the design of toolkits. In the next section, we will use this literature review as a basis to compile a list of requirements for toolkits.

We used the earliest and widely-cited key studies by Veale et al. [126] and Holstein et al. [65] as the base for a snowball sampling literature review. Both are interview studies, where practitioners (in case of Veale et al. public sector decision-makers and in the case of Holstein et al. ML developers) are interviewed to understand what tools, additional research and organizational reforms they need to conduct better fairness audits. Within all papers that cited either of two studies, we used the search query "interview study + fairness assessment" to extract similar papers in which practitioners are interviewed to understand their practices and need for assistance in conducting fairness audits.

Based on the results we identified two relevant lines of research. The first are interview studies, where people working with ADM and other ML systems reported their current practices and obstacles when assessing or ensuring the fairness of these [35, 65, 126]. These studies do not solely focus on the potential of interactive toolkits in addressing these problems but also explore reformations in the organizational and legal sphere.

More recent studies also directly investigated the potential of fairness toolkits, in facilitating bias audits. Here, possible auditors were interviewed or asked to test tools, in order to identify their requirements in them [84, 86, 110].

In section 1.3.1 and 1.3.2 we will summarize both lines of research and identify the design considerations (DC) for interactive auditing tools that emerge from them. In doing so we only focus on the considerations for tools that help in detecting bias in ADM systems (and not other ML applications). Further, we will only concentrate on the functionality and not the usability of such tools. As we will see there is a lot of overlap in the design considerations that have become apparent from the different interview studies.

1.3.1 Exploring current practices for bias detection in ADMs

The earliest significant study on algorithmic bias audits, conducted by Veale et al. [126], involved semi-structured interviews with 27 individuals from the public sector. Semi-structured interviews are a research method where participants are asked a series of predetermined open-ended questions, but the interviewer also has the flexibility to ask additional follow-up questions to explore topics in more depth. In this study, the interviewees, who utilized ADM models for decision-making in areas like taxation or policing, were asked to share their experiences with the models, express fairness concerns, and discuss obstacles they encountered. While many of the reported issues lay on an organizational level, interviewees also revealed some practices and concerns that are relevant to the design of auditing toolkits: they reported that they were aware of discriminatory effects of ADM models and that they, therefore, avoided the use of sensitive attributes when building such models. Further, they were wary of utilizing variables like "home location" in their model, as they might serve as a proxy for the sensitive attribute "race". Still, guidelines of which variables to avoid in models were more of an informal nature, as this is dependent on the decision task. Although not directly mentioned by Veale et al., the fact that practitioners do not directly use sensitive attributes in their models but have no formal way of identifying all proxy attributes poses serious considerations to the design of auditing toolkits: first, if sensitive data is available but not being used, tools should provide the functionality to detect proxy attributes, based on their correlation with sensitive attributes (**DC_identifying_proxies**) Second, if no sensitive attributes are available, tools should still allow auditors to conduct a fairness analysis. We will refer to this design consideration as **DC_no_sensitive_attributes**.

One year after the interview study by Vaele et al., Holstein et al. released another study building on top of their results. They conducted 35 semi-structured interviews with ML developers, to find out about the obstacles they experience when assessing and improving the fairness of ML systems. After the interviews, they also conducted a survey to see whether their results were generalizable to a wider public. Similarly, as in Veale et al. many of the identified issues lay on an organizational level or were specific to ML applications that are not the focus of our paper. Still, practitioners also reported technical issues in assessing/improving the fairness of ADM models, which should be considered when designing auditing toolkits. The first issue relates to the already discussed design consideration **DC_no_sensitive_attributes**, as practitioners reported that often access to sensitive attributes is lacking. Another main issue relates to the preferred intervention stage when improving the fairness of a system. Practitioners revealed that when a model appears to be biased, they inspect the training data this model was based on, to think about ways in which collecting more data or pre-processing the data can help in mitigating the bias. Relating this to the design of interactive auditing toolkits, this means that tools should help auditors in inspecting the training data so that they can identify causes of prediction biases and resolve them. We will name this design consideration **DC_identifying_bias_causes**. Further, auditors mentioned that the closer inspection of input data is also important for assessing the quality of the test set that a model is audited on. After all, only if the test set is representative of the data that the model is applied on, the results of the fairness audit can be generalized. We will refer to this design consideration as **DC_fair_testset_design**.

Another design consideration we extracted from their work relates to practitioners' fear that a wide range of biases may creep into a model and identifying all of them is time intensive. Hence, they do not want to waste efforts on identifying occasional "one-off" mistakes from a model but want to prioritize big, systemic biases, that are unlikely due to chance. As we will see in section 1.4.3 there are ways in which auditing toolkits can meet this requirement (**DC_prioritize_systemic_biases**). Connected to this, practitioners still fear, that they have blind spots in analysing the fairness of a system and that they do not think of all the attributes that can serve as grounds for discrimination. Hence, creating a tool that can automatically suggest possibly discriminated subgroups or individuals, could be a way to accommodate this fear (**DC_account_for_blindspots**, described in further detail in section 1.4.3.2)

The final paper we are going to discuss was written by Constanza-Chock et al. [35]. They interviewed 10 different auditors of ADM systems, that were either researchers, CEOs of dedicated auditing companies, or leads of internal company teams responsible for bias audits. Their goal was to identify current auditing practices, as well as obstacles on organizational, technical and legal levels these auditors faced.

One interesting discovery was that many auditors currently favour custom-built toolkits over standardized ones. The preference for custom solutions is attributed to the fact that standardized toolkits may not always fit tailored use cases, and some interviewees expressed concerns about the overemphasis on quantitative measures of fairness that try to express unfairness in a single number, without further consideration of its context or origins. However, despite this preference for customization, we believe there are compelling reasons to enhance standardized toolkits, as they are readily available to the public, more cost-effective, and can be swiftly implemented, unlike developing a new custom tool from scratch. While the study conducted by Constanza-Chock et al. did not directly address how to capitalize on this potential, we identify several ways to overcome the established disadvantages. First, to address the applicability to tailored use cases, toolkits should offer a wide range of functionalities that cater to various scenarios. Additionally, allowing some degree of customizability would be beneficial, enabling users to not only examine pre-defined fairness or performance measures but also define their own metrics (**DC_variability_and_customizability_of_metrics**). Related to the second concern, that tools express the fairness of a system only through quantitative measures, it is essential that tools also encourage deeper analysis of biases: for instance, by letting users inspect the training data behind a model, tools can allow users to reason about the causes of a models' unfairness (**DC_identifying_bias_causes**). Further, by letting users not just inspect demographic subgroups (e.g., based on gender or race) but also subgroups based on other attributes in the data, users are encouraged to contextualize biases better and understand their occurrences (**DC_bias_contextualization**).

Another interesting finding that reveals a design consideration for interactive tools is how auditors currently deal with intersectional bias analysis: Constanza-Chock et al. found that auditors generally have the intent to perform such analyses, but in practice could not provide many cases in which they were conducted. They hypothesised that this was likely due to the general difficulties surrounding such analyses, like dealing with a large number of small subgroups and not being able to identify all marginalized groups. Despite such difficulties, the importance of identifying and understanding intersectional biases is clear, which is why tools should support and facilitate such analysis (**DC_intersectional_analysis**).

1.3.2 Exploring the potential of tools

The studies discussed in the previous section address auditors' current practices and concerns when assessing the fairness of ADM systems. In this section, we will examine the studies that explore how practitioners think toolkits can help in this assessment. The first of these studies was conducted by Law & Du. They held 10 semi-structured interviews with ML practitioners of the same company, working on different projects. They introduced the practitioners to the case example of bias detection in the *Adult dataset* (the same dataset we have described in section 1.2 of this paper) and then asked them about their encounters with bias detection in ADM models and how they thought tools could help them. As we already identified as **DC_no_sensitive_attributes**, practitioners reported that not having access to sensitive attributes was a major obstacle for them in auditing a systems' bias and that having tools that can help with that would be highly useful. One concrete suggestion was to make tools that automatically predict the sensitive attributes of data instances, based on other features in the input data. Similarly as found by Holstein et. al, interviewees also expressed their fear of bias audits becoming unscalable and suggested implementing functionality in tools that allow them to prioritize big, systemic biases (**DC_prioritize_systemic_biases**) and functionality to ensure they do not miss any of them **DC_account_for_blindspots**). Finally, they also mentioned the importance of having tools that allow them to assess the training data, to identify the causes of a model's biases (**DC_identifying_ bias_causes**).

Richardson et al. conducted another study investigating toolkits' potential in facilitating algorithmic bias audits. In a usability study, they let 20 ML practitioners test one of two fairness toolkits (Aequitas [115] or Google Fairness Indicators [122]) and let them reflect on the usefulness of these. They used the results of this study to set up a rubric with tool requirements. While this rubric also contains points regarding the tools' usability and tools that could be used for a broad range of ML applications, we will concentrate on the functional requirements related to bias detection in ADM systems. All of these requirements are related to design considerations we already established from previous literature: **DC_variability_and_customizability_of_metrics**, **DC_intersectional_analysis**, **DC_no_sensitive_attributes**, **DC_bias_contextualization** and **DC_identifying_ bias_causes**.

Another relevant paper is the work by Lee & Singh, who conducted semi-structured interviews with ML practitioners to review programming libraries like IBM Fairness 360 or Fairlearn [13,95] that provide pre-defined metrics and algorithms to analyse and mitigate the bias of ADM systems. Based on the interviews, Lee & Singh establish how these libraries could be improved. The first design considerations concern the need to inspect the training data to identify bias causes and find possible proxies for protected attributes (**DC_identifying_bias_causes**, **DC_identifying_proxies**) Second, they were concerned about the customizability of tools to their use cases **DC_variability_and_customizability _of_metrics** and the degree to which they could handle more complex form of bias (e.g., bias based on multiple non-binary sensitive attributes, **DC_intersectional_analysis**). Third, they expressed their interest in tools that highlight the significance of biases, so that they would not waste time inspecting disparities that are due to random chance **DC_prioritize_systemic_biases**.

Even more recently a study was conducted by Nakao et al., who developed a new prototype for an interactive auditing toolkit after they conducted several workshops to identify stakeholders' needs in such tool [98]. They specifically focused on fairness audits in the context of a loan allocation system and therefore interviewed both data scientists and loan officers as potential auditors of this system (note that we do not review their developed prototype as part of our fairness tools since it is not publicly available). In their study they identified the following design considerations: **DC_variability_and_customizability_of_metrics**, **DC_intersectional_analysis**, **DC_identifying_proxies** and **DC_bias_contextualization**.

1.4 Functional Requirements for Tools

1.4.1 Functionality for detecting bias in models' predictions

The first main category of requirements regards a tool's functionality of letting users identify bias in a model's predictions. In the following sections, we will cover what forms of biases should be detectable by toolkits (section 1.4.1.1), how it is important that intersectional bias can be analysed (section 1.4.1.2) and how tools should offer bias analysis that goes beyond sensitive attributes, in case that there are proxy attributes in a model or attributes that make the treatment of groups with different sensitive attribute values justifiable (section 1.4.1.3). In each section, we will also discuss to which extent our reviewed tools offer the required functionality.

Table 1.1: Requirements related to tools' functionality to let auditors find bias in an ADM model's predictions.

Functionality for detecting bias in a model's predictions	
Different forms of bias can be detected **[DC_variability_and_customizability_of_metrics]**	
Some standard bias measures are supported:	
Outcome based (group)	AE DL FS (RD) (WI)
Actual vs. Outcome based (group)	AE FS FV RD WI
Probability based (group)	FS (RD) (WI)
Similarity based (individual)	FS RD WI
Causal based	/
Tool provides customizable bias metrics	/
Intersectional bias can be explored **[DC_intersectional_analysis]**	
Bias based on non-binary sensitive attributes	AE FV RD WI
Bias based on multiple sensitive attributes	FV RD (WI)
Prediction bias beyond sensitive attribute(s) **[DC_bias_contextualization], [DC_identifying_proxies],** **[DC_no_sensitive_attributes]**	
Tool lets user contextualize differences in outcomes	DL RD WI
Indirect Bias Analysis (**with** access to sensitive attributes)	
Functionality to find proxies	FS RD WI
Functionality to relate proxies to decision attribute	see section 1.4.1.2
Indirect Bias Analysis (**without** access to sensitive attributes)	
Estimate sensitive attributes from data	/
Functionality to relate (possible) proxies to decision attribute	see section 1.4.1.2

1.4.1.1 Different forms of bias can be detected

This requirement pertains to the design consideration **DC_variability_and_customizability_of_metrics**, which has been established by reviewing the works of [35, 110]. To reiterate, some practitioners have avoided using toolkits because they do not provide implementations for all the bias metrics relevant to their decision task. Therefore, an implementable solution is to create toolkits that offer a wide range of standard metrics, that can further be customized to their specific use case.

Some standard bias measures are supported Because bias is such a complex and non-arbitrary concept, various, often incompatible definitions can be used. While it may be tempting to only choose one of these definitions and assess a system's fairness accordingly, researchers have warned against such simplifications [11, 118]. Take for instance the measure of "Equal Opportunity" in the example of our loan allocation system. According to this definition, a system is free of bias if the "true positive rates" among all groups of interest (e.g. all genders) are equal. The "true positive rate" is defined as the probability that a loan applicant for which an ML model predicted a positive outcome (i.e. being approved for a loan) also has a positive label in the data. While at first sight, this definition may sound "fair" it does not account for the "unfairness" that might be present in the labels, based on which a model's errors are assessed. In the case of the *Adult dataset*, it could be argued that the inequality in high and low incomes between genders reflects the gender pay gap that is a result of direct discrimination as well as unequal (and possibly unfair) societal expectations and opportunities for different genders [26]. Thus, the fact that more men than women have a positive label (i.e. high income) in this data does not mean that this bias should be replicated by an ML model trained on it. Especially, if the labels would be used as a proxy for who deserves a loan and not, it can easily be argued that consistently giving more men than women a loan, would only increase existing gender inequalities. To account for existing biases in the labels, it is possible to choose a "bias-transforming" fairness goal [128] like "Demographic Parity". With this measure, we ensure that an equal portion of men and women are granted a loan by the system. Still, also this measure comes with disadvantages. For instance, it ignores differences in qualifications or eligibility of population groups that could justify a difference in outcome (i.e. loan vs no loan) [11]. One way to address this problem is by focusing on similarity-based fairness measures, that are based on the principle of "treating likes alike": individuals that are similar in terms of their eligibility for a loan should obtain the same outcome [11].

As has become clear from this example, there is a variety of bias measures to take into account when auditing an ML system and there is no single criterion that "makes or breaks" the fairness of a system. Hence, a tool must support a wide range of these metrics so that auditors can choose one or multiple to inspect based on the given use case. To distinguish between the different forms of fairness that should be measurable with a toolkit we will, similarly as [109], make use of the five bias categories specified by [127].

Group-Based Measures The three definitions falling under the subcategory of *group-based bias measures* measure whether there are substantial differences in treatment between two or more groups (e.g., men vs. women vs. non-binary). This can first be measured by comparing the classifier's outcomes on the groups, second, by comparing the classifier's errors on them, and third by comparing the classifier's predicted probabilities on them. As can be seen in Table 1.1, most auditing tools support the bias definitions based on classifiers' errors, with Equal Opportunity (one of the measures discussed previously) being one example of such definition. We have already pointed out how these error-based measures are not appropriate to account for the bias present in the ground-truth labels. Hence, the fact that many tools do not support outcome-based measures (that do account for this bias) poses a serious shortcoming for their applicability. Additionally, the fact that so few tools support probability-based measures, is another drawback. Most ADM models do not directly output binary decision labels for a prediction task, but instead prediction probabilities, which can then be translated to binary labels by applying some decision threshold on them. Yet, this threshold is quite flexible and may even change throughout a system's deployment. For instance, in the case of our loan approval system, this may depend on the bank's resources and the number of loans it can grant [72]. To be able to guarantee the fairness of a system, independent of a chosen decision threshold, it is thus useful to have tools that allow for probability-based bias assessments [72]. One example of such measure is the "Balance for positive class", demanding that for all instances with a positive decision label in the data, the average prediction probability is the same across groups [80] (i.e. the average prediction probability for women with a positive label is the same as the average probability for men with a positive label). Satisfying such a goal gives some guarantee that a model's prediction will still be fair once the decision threshold changes.

Currently, only three tools partly allow for the inspection of probability-based bias measures. FairSight operates on the ranking produced by a classifier, which is obtained when ordering the decision instances according to their prediction probabilities. The tool then prompts the user to specify which top-k instances of this ranking will be granted a loan and calculates various bias metrics based on the protected/unprotected individuals represented in this ranking (see Figure 1.4 (B)). While this gives some insights into the probability-based fairness of the corresponding model, the tool does not operate directly on prediction probabilities but only on the obtained ranking. The other two tools that partly allow the exploration of probability-based bias measures are the WhatIf tool and the RAI dashboard. Both allow users to visualize the prediction probabilities for different subgroups, as can be seen in Figure 1.10 (A) for the RAI dashboard and Figure 1.7 (A) for the WhatIf tool. However, neither of these visualizations is accompanied by formal measures (note, that the same holds for output-based bias measures: both tools allow for visualization of them but do not provide exact measures). Adding formal measures would thus be an easy way to improve the suitability of both tools.

Adding a wider variety of bias measures to a tool like DiscriLens, which has been developed with one specific bias metric in mind (in DiscriLens' case "conditional demographic parity"), will prove a bigger challenge. This highlights the need to take a broad perspective when designing auditing tools and make them flexible for different tasks and fairness notions.

Similarity based Measures The fourth group of bias definitions, in addition to the three described above, are similarity-based ones, which define bias s on an individual level by comparing the outcome of a data instance with those of similar ones. Currently, only FairSight, the RAI dashboard and the WhatIf tool support users in exploring these definitions. FairSight does so by enabling the user to conduct a "nearest neighbour" analysis, where the user can select an instance and compare their predicted ranking position to those of similar ones (see Figure 1.5 (d)). The WhatIf tool and the RAI dashboard on the other hand provide a "What-if" analysis. Here users can change feature values of an instance of choice, and observe how this affects the prediction. To illustrate, look at Figure 1.7 (B), showing this component of the WhatIf tool. Here a user has selected a female data instance and changed their sex to "male". The outcome of this change is that the probability of granting a loan raises from 0.283 to 0.535. There are two primary reasons why this kind of similarity-based fairness analysis should be an essential component of any auditing tool:

- First, consider our loan approval system and imagine that it satisfies the outcome-based fairness measure demographic parity in regard to gender (i.e. for all gender groups it grants the same proportion of loans). While this system may look fair on the outside, there is no guarantee that the people who get granted a loan also are eligible for one, or that the reasoning behind granting a loan is fair. It may for instance still behave in an individually discriminating way like in the example given above
- Second, in the study by Richardson et al., interviewees mentioned how they found it easier to understand global patterns of discrimination (as given by group-based fairness definitions) when being provided with individual examples of discriminated instances. This finding is also backed up by [41].

Both reasons should provide developers of new tools with motivation to include similarity-based analyses in them. Still, it should be noted that similarity-based measures also come with disadvantages, the biggest one being that it is not clear how to define similarity and how similar two instances should be to receive the same decision outcome. In the case of a loan allocation system, it is e.g. clear that a man and woman who are identical in all features except their sex, should not be treated differently. However, if a man and woman also differ on a relevant attribute (e.g., their current employment status), this is not so arbitrary. On the one hand, a difference in this attribute can justify handing out a loan to one person but not the other. On the other hand, differences in these attributes may reflect systemic and societal gender inequalities (e.g., different lengths of parental leaves, women working more part-time, etc.) that an auditor needs to account for [67]. Hence, for functionalities like the "nearest neighbour analysis" in FairSight, it can be useful if a tool lets auditors define their own similarity metric to allow for such considerations. We will further elaborate on this point in section 1.4.1.1.

Causal-based Measures The fifth and last category of bias definitions are causal-based ones [127]. These definitions are the most distinct, as they do not solely define bias on the predictions of a model but also on the causal relationships that are assumed to underlie it. In other words, we use causal-based definitions to examine whether there are discriminatory causal relationships between a sensitive attribute and a decision attribute in a model's decision-making. Currently, no tool allows the user to investigate these causal notions. As causal bias definitions lay in a niche research area within the fairness literature, it may not be surprising that no tools support their exploration. Still, it should be noted that there may be great potential in incorporating them into interactive tools, especially by visualizing the causal relationships within a model through causal networks. The paper by [27] gives a good overview of how causal networks could help in the detection of bias in an ML model. Additionally, [139] and [98] present tools through which causal analyses can be conducted, and also point out the merits of adopting a causal framework. To give a more concrete example, refer to Figure 1.12 displaying a

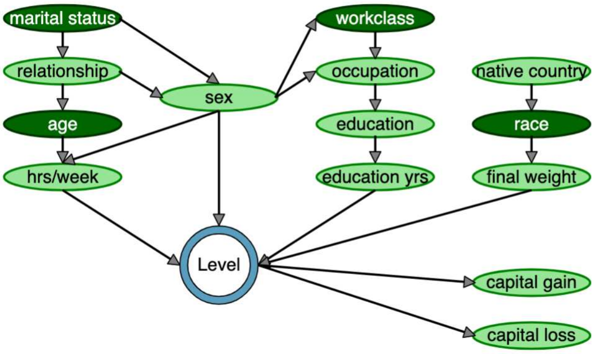


Figure 1.12: A graph visualizing the causal relationships within an ADM model trained on the *Adult dataset*. This graph is visualized as a part of a auditing tool developed by [139]. Though this tool is not openly accessible, its design and its use of causal fairness definitions can still serve as an inspiration for other toolkits.

graph from [139], that visualizes the causal relationships within the *Adult dataset*. In their tool, [139] show this graph to let users reason about problematic relationships between attributes like "sex" or "race" and the decision attribute "Level" (short for "level of income"). In this case, we see that peoples' sex has a direct causal effect on their income, but is also linked to other attributes (e.g., the number of working hours) that may influence income levels. The cited papers give more information on how to quantify these relationships and how auditors could use visualizations like these to reason about the biases within a system. For instance in this case it is clear that utilizing the ADM model based on the causal relationships in 1.12 is problematic, since in this model there is a direct link between "sex" and "income". If the causal relationships within the model were different, and there was only an indirect link between people's sex and their income level (e.g., explained by the link to different working hours between different sexes), auditors could apply different reasoning as to why this link may or may not be acceptable.

Though the tools by [139] and [98] are not openly accessible, their design may still serve as an inspiration to add further functionality to existing tools.

Customizable metrics While it is useful if a tool provides a couple of standard bias measures by default, our reviewed interview studies revealed that auditors would also like toolkits in which they can customize their own metrics (see **DC_variability_and_customizability_of_metrics** in section 1.3) [35, 86, 110]. This also relates to findings in other ML literature, where practitioners explain how they evaluate their products on organization-defined and product-specific metrics, rather than standard ones [120]. While the customizability of metrics is arguably a broad requirement, we already touched upon some ways in which this requirement could be fulfilled, like allowing users to define a similarity function for similarity-based fairness measures (see section 1.4.1.1) or allowing them to specify a decision threshold to translate prediction probabilities to prediction labels (see section 1.4.1.1). Another suggestion that came forth from the interview studies discussed earlier is to let users define metrics that can be used to assess a model's fairness in non-binary prediction tasks, like multi-class problems or regression problems. In the case of a loan approval system, it might, e.g., be of interest not just whether an individual gets granted a loan but also what the height of that loan is, and whether that is equally distributed among demographic groups.

1.4.1.2 Intersectional bias can be explored

Another design consideration for the development of tools is their functionality to detect intersectional biases (**DC_intersectional_analysis**). "Intersectional bias" is a term that was 1989 coined by Kimberly Crenshaw, to describe the discrimination that black women faced in employment that could neither be fully explained by discrimination against sex, nor discrimination against race [36]. Since then, the term has been used to describe how people who come from a combination of marginalized groups (based on gender, race, religion, class, and other identity markers), face different levels of discrimination than cannot be explained by the "sum" of discrimination faced by each marginalized group in isolation. ADM systems may also behave in intersectionally discriminatory ways, which is why tools should assist in the detection of those.

To facilitate this, there are two functional requirements a tool should fulfil: first, it should allow the analysis of bias based on non-binary sensitive attributes, and second, it should allow the analysis based on combinations of these attributes. Both points are elaborated on in the next paragraphs.

Bias based on non-binary sensitive attribute(s) The first consideration that needs to be made when conducting an intersectional bias analysis, or even when analysing bias from a single-axis, is which identities to include per sensitive attribute [130]. As sensitive attributes are typically non-binary, auditing toolkits must support this non-binary analysis. Out of our six tools, all do so except DiscriLens and FairSight. The risk of using such simplifications should not be underestimated. Take for instance the attribute "race"; using a tool like DiscriLens or FairSight, we are forced to discretize this feature into two groups, e.g. "white" and "non-white". Yet, for any domain expert using such tool, it is clear that this discretization does not account for all the different types and levels of bias different non-white racial groups may face [65, 130]. Looking for instance at Figure 1.10 (A), we see the distribution of predicted probabilities for the group of white men, coloured men, and black men. The model predicts higher probabilities of granting a loan for the group of white than coloured men. However, the difference in prediction probabilities (and also False Negative Rates) is even larger, when comparing the group of white and black men. This indicates that within the group of coloured men, black men face especially averse effects. If an auditor would use a tool that only allows bias detection on binary-sensitive attributes, they would miss this important pattern. Fortunately, it should not be too difficult to allow for non-binary bias analysis in DiscriLens and FairSight. Both tools heavily rely on colours to encode different groups of interest in their data visualization/exploration. Adding more colour options to the tools is one possible way to allow for fairness analysis of non-binary sensitive attributes.

Finally, note that in the question of which categories to include per sensitive attribute, also broader issues need to be addressed, for instance how attributes like race or gender were recorded (i.e. are they self-reported or recorded by the data collectors?). While it is not possible for a tool to address these issues on a technical level, they can still pose serious threats to the fairness of an ADM system and therefore should not be ignored in an audit [40, 130].

Bias based on multiple sensitive attributes Once it is clear which categories to include for each sensitive attribute, the next step for an intersectional analysis is to decide on the combinations of attributes that need to be inspected. To be able to conduct such analysis with an interactive tool, the tool must support the fairness analysis based on multiple sensitive attributes. Out of all tools, only FairVis and the RAI dashboard fully do so. Indeed, when using this functionality we see that a model trained on the *Adult dataset* also displays signs of intersectional discrimination. Looking at Figure 1.2 (b) we see the "False Negative Rates" for different subgroups based on people's "sex" and "race". We observe that this rate is already quite high for women, even higher for black people and highest for black women. This knowledge is crucial for a fairness auditor to decide on how to improve an ML system. In this case, an auditor could e.g. recommend that before the system can be deployed more data needs to be gathered for this subgroup. If an auditor would only analyse one sensitive attribute at a time, they might not have found this solution, and might only suggest collecting additional data for women and black people, rather than the intersection of both. Following this example, more tools must allow the analysis of intersectional discrimination. The WhatIf tool already partly supports this feature, but only for subgroup combinations based on two sensitive attributes. Still, the way this tool as well as FairVis and RAI dashboard allow for intersectional bias analysis can serve as inspiration for other tools: the functionality works by letting users generate subgroups of choice (see e.g. Figure 1.2 (A) for FairVis, Figure 1.8 (A) for the WhatIf tool and Figure 1.9 (B) for the RAI dashboard), and then inspect and compare all fairness metrics across all user-generated groups. Similar functionality could be added to other tools.

1.4.1.3 Prediction bias beyond sensitive attributes

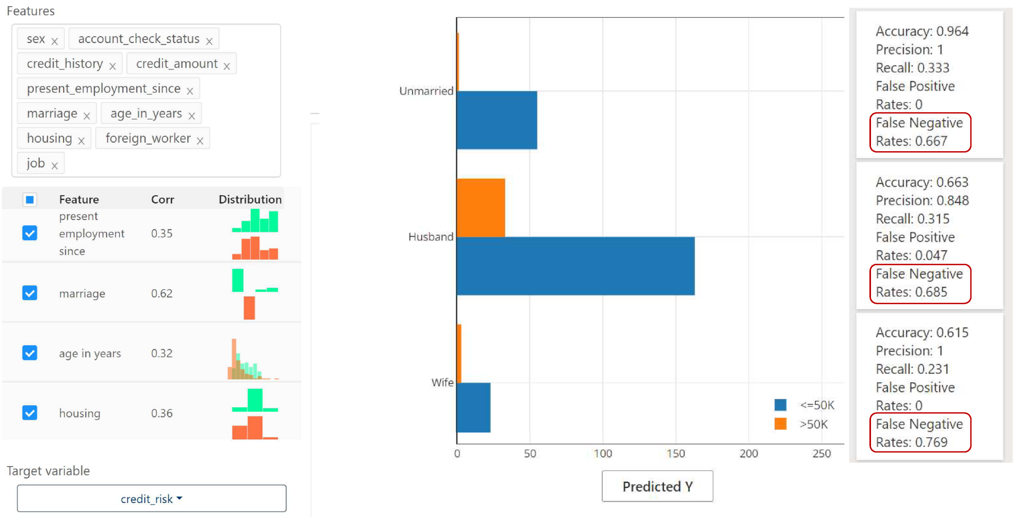
In the previous section, we assumed that in the fairness assessment of an ADM model auditors have access to all relevant sensitive attributes and that they are only interested to observe disparate behaviour of a model based on these attributes. The interview studies revealed, however, that current auditing practices often go beyond the analysis of just sensitive attributes for two reasons: first, it is important to contextualize differences in outcomes between different demographic/sensitive groups, since a model's decision to treat groups differently may be justifiable (**DC_contextualize_biases**, [35, 98]). Second, discriminatory biases may not always be based on attributes that are legally protected (e.g., gender or race) but on attributes that might serve as a proxy for these (e.g., zipcode for race), a phenomenon known as indirect bias. Functionality to conduct an indirect bias analysis, both in the case in which auditors have access to sensitive attributes and those in which they don't is, therefore, essential in a toolkit as earlier indicated by **DC_identifying_proxies** and **DC_no_sensitive_attributes** [65, 84, 86, 110, 126].

Contextualize differences in outcomes The functionality to contextualize biases is important to understand why an ADM model may make less preferable decisions for some population group over another. In some cases, a difference in treatment may be justifiable by so-called "explainable attributes" [75]. When for instance in our use case a classifier decides that more men than women should receive a loan, this is not necessarily problematic if this can, e.g., be explained by women in the data working in lower job positions than men, indicating that they have less financial means to pay back a loan. DiscriLens is a tool specially developed to contextualize biases and understand if they are explainable. Here users specify a list of explainable attributes, and the tool automatically highlights the cases where a protected and an unprotected group have a high difference in positive decision probability, conditioned on these attributes. In other words, it only displays biases that are not explainable. For instance, in Figure 1.3 (E), we see that when specifically filtering for people with high education levels and high amount of workinghours, still more men than women get granted a loan. Similarly, the "Model Statistics" component of the RAI dashboard allows for the analysis of non-explainable discrimination, by visualizing the prediction outcomes for the group of highly educated men and women, to see if there are fundamental differences in both (see Figure 1.10). Note how powerful the "subgroup generation" functionality is in the RAI dashboard, as the same mechanism can be used to study intersectional discrimination (see the previous section). Thus, adding similar functionality to other tools should make them more suitable to auditors' needs.

One final note for the contextualization of bias, is that the choice of "explainable" attributes should always be carefully considered by a domain expert. In the previously mentioned example, of women less likely to receive a loan because of having lower job positions than men, an expert should always consider the question of why this is the case and whether this is the result of historical bias (which in our example might very well be the case, given that women are known to not receive the same job opportunities as men). To make up for this already existing bias, it may not make sense to ignore "explainable" patterns of discrimination, but instead, critically question the extent to which "explainable" discrimination is legitimately explainable [128].

Indirect Bias Analysis (with access to sensitive attributes) As explained earlier, indirect discrimination in ADMs occurs when a model does not directly make use of sensitive attribute information to derive its decisions, but when it relies on attributes that are proxies for these. One famous example is the practice of redlining, where a model indirectly disadvantages racial groups, by using the zip code of people as a factor in its decisions. As we have found in our review, auditors are highly aware of the phenomenon of indirect discrimination, which is why they require tools that allow them to analyse it. In the case that they have access to sensitive attributes, like gender or race, an auditing tool can facilitate this analysis by first allowing them to identify proxy attributes and then letting them explore a model's behaviour on these. FairSight helps users in the first step, by providing visualizations of how attributes are differently distributed among sensitive groups, as well as giving a correlation measure between them (see Figure 1.4 (A) to inspect the component for this tool, and Figure 1.13a for a specific case example). In Figure 1.13a we see that the feature "marriage" is considerably differently distributed between the protected and unprotected group, caused by the fact that "Wife" is a feature value that is only applicable to women, while "Husband" is a value only applicable to men. The difference in feature distribution is also indicated by a high correlation measure between the feature "marriage" and sensitive attribute "sex". To allow for the detection of proxy attributes, also the RAI dashboard, the WhatIf tool and FairVis enable users to visualize how attributes are differently distributed for population groups (see Figure

1.10 (B) for the RAI dashboard, Figure 1.7 (A) for the WhatIf tool and Figure 1.2 (A) for FairVis). However, these visualizations are not accompanied by correlational measures. Concerning the second step in the identification of indirect bias, i.e. the exploration of a model's predictions on the different values of this variable, only the RAI dashboard, the WhatIf tool and FairVis allow doing so (this step essentially boils down to inspecting a models' behaviour on a non-binary sensitive attribute (see section 1.4.1.2)). After e.g. having found that an individual's relationship status is highly correlated to their sex, we could use the RAI dashboard to visualize a model's performance on different population groups determined by this proxy attribute. In Figure 1.13b we indeed see that the model performs unfairly on the group of "Wives", making more false negative errors for them than other "relationship" groups.



Features

sex × account_check_status × credit_history × credit_amount × present_employment_since × marriage × age_in_years × housing × foreign_worker × job ×

☐	Feature	Corr	Distribution
☑	present employment since	0.35	
☑	marriage	0.62	
☑	age in years	0.32	
☑	housing	0.36	

Target variable

credit_risk ▾

Unmarried: Accuracy: 0.964 Precision: 1 Recall: 0.333 False Positive Rates: 0 False Negative Rates: 0.667

Husband: Accuracy: 0.663 Precision: 0.848 Recall: 0.315 False Positive Rates: 0.047 False Negative Rates: 0.685

Wife: Accuracy: 0.615 Precision: 1 Recall: 0.231 False Positive Rates: 0 False Negative Rates: 0.769

(a) FairSight's feature to detect proxy attributes

(b) The RAI dashboard lets users inspect differences in prediction outcomes for redlining attribute values

Figure 1.13: Functionality to (a) find proxy attributes that are highly correlated to sensitive attributes and to (b) detect the relationship between these proxy attributes and a class label

Indirect Bias Analysis (without access to sensitive attributes) Analysing the occurrence of indirect discrimination is complicated considerably when the training data of a model does not contain any "traditional" sensitive attributes but may contain proxies for these, which in turn cannot be directly identified as such. As we have seen in section 1.3 this is a very real concern among auditors: often sensitive information is not collected for a decision task (since it may be even illegal to do so), yet without this information, it is hard to identify possible disparate impacts of a model for different sensitive groups [35, 65, 84, 110, 126]. In the interview studies by [65, 84] auditors made some suggestions on how this concern could be addressed on a technical level, using interactive tools: they suggest making tools that can estimate sensitive attribute values for each individual based on the rest of their information. To illustrate, some ML practitioners interviewed by [65], already developed systems that use information about peoples' IP addresses (disclosing information about their home location) and names to estimate their sex and ethnicity. Still, they were wary about additional biases introduced by this process and also had concerns about storing (inferred) demographic information and the associated risk of data leakage or misusing this data for secondary purposes [65]. Currently, none of our reviewed tools supports the estimation of sensitive information based on other attributes in the data. However, given the risks of this approach, it is also questionable to which extent this feature is desirable to deal with the analysis of indirect bias.

Other literature on fairness in ADM explores alternative ways to deal with this problem. First, though of less interest in our paper, there are legal regulations that could be enforced, to only allow trusted third parties to access sensitive attributes solely for auditing purposes [78, 124, 125].

Another more technical approach for unravelling patterns of indirect discrimination is the exploratory analysis of the ADM model [125]. In a 2017 paper Veale & Binns suggest that exploratory analyses could be used to find interesting patterns in the data, that could afterwards be more closely inspected for possible correlations with sensitive/demographic groups (using additional data sources) [125]. This approach to identifying indirect discrimination was also suggested by Ruggieri et al., who extracted potentially discriminatory association patterns from the data (e.g. *IF "zipcode" = XYZ THEN "no loan"*) to then use additional databases to find whether the premises of these rules (i.e. *"zipcode" = XYZ*) relates to sensitive information of individuals [114]. While additional resources are needed to perform this second step, tools can facilitate the execution of the first step by enabling auditors to analyze performance/fairness measures based on the value/value combinations of other attributes in the data. This requirement was already established in section 1.4.1.2 and further explored in section 1.4.1.3, highlighting how important a flexible design of toolkits is and how limiting it is if tools only support the fairness analysis based on one, pre-determined sensitive attribute.

1.4.2 Functionality for detecting bias in models' input data

The rubric given in Table 1.1 focuses on a tool's functionality to find biases in the predictions of a model. However, as we have found through our literature review, auditors also find it important to inspect the input data for possible biases for two reasons: first, to understand where biases in a model's predictions may be coming from (**DC_identifying_bias_causes** [35, 65, 84, 86, 110]) and second, to ensure that the fairness audit of the model is conducted on a representative and "fair" test set (**DC_fair_test_set_design** [65]). Lastly, also note that in our paper we focus on assessing the input data as part of the audit *after* model development. Still, the assessment of the training data should be an essential step *before* a model is trained, as basing a model on highly biased data might not be desirable in the first place. Of course, the requirements listed in the upcoming section still hold for tools that would be used for this purpose.

1.4.2.1 Finding bias causes in training data

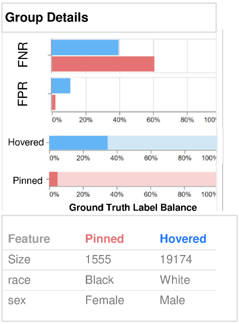
In discussing a tool's functionality to find bias causes we will distinguish between the different bias causes established by Suresh & Guttag [121]. Note that many bias causes

Table 1.2: Requirements related to tools' functionality to let auditors find biases in an ADM model's input data.

Functionality for detecting bias in models' input data **[DC_identifying_bias_causes], [DC_fair_test_set_design]**	
For identification of bias causes in the training data, let user...	
Inspect the relation between attributes and ground truth	FV RD WI
Inspect relation between features	FS FV RD WI
Compare train and test set	/
For identification of biased pattern in test data, let user...	
Inspect subgroup sizes of interest	FV RD (WI)

(e.g., errors in how the data was collected) are not completely identifiable on a technical level and that we will merely focus on those causes whose identification can be facilitated by auditing toolkits.

Inspecting the relation between attributes and ground truth One possible source of bias in ADM models comes from the bias that may be present in the ground truth that the model is trained on, that the decisions that were made historically for the decision subjects (e.g., high vs. low income or loan vs. no loan) in the training data. The ground truth can be subject to *historical bias* or *measurement bias* [121]. The former occurs when the label has been recorded correctly, but contains patterns of historical inequalities between population groups (e.g., women being recorded to have lower income than men). The latter (i.e. measurement bias) occurs when due to errors or biases in the decision process, individuals did not get the label they were eligible for (e.g., women who are not granted a loan, even though they would have paid it back if given the opportunity). While it may not be possible to distinguish between these forms of bias in the labels, it is still crucial that tools allow auditors to inspect the labels, to understand if they are favoured more towards some groups than others. For this, tools should support all purely label-based fairness measures, as discussed in section 1.4.1.1 (i.e. the measures that are not based on prediction errors/prediction probabilities) on the model's ground truth: outcome-based measures, like demographic parity, similarity-based measures and causal measures. When tools provide these measures, the same requirements hold as discussed in section 1.4.1.2 and 1.4.1.3: the measures should be applicable on intersectional groups and the bias measures should go beyond sensitive attributes, to contextualize biases and understand patterns of indirect discrimination. Currently, FairVis, the RAI dashboard and the WhatIf tool allow for partial analysis of bias in the ground truth labels. In Figure 1.14, an example is shown where predictions of an ADM model are biased against black women (high False Negative Rates) and favoured towards white men (higher False Positive Rates). When we inspect the ground truth label balance, we find a reason for why the ADM model is more biased towards predicting positive labels for white men: they have more than double the ratio of positive labels in the ground truth than black women. Though this information is useful to understand where the model's bias comes from, it would (among others) be useful if FairVis would let users contextualize this bias, to understand whether the difference in ground truth label balance is "explainable" by other attributes. Note, how the functionality to contextualize biases is lacking both in FairVis' functionality to assess fairness in predictions as well as fairness in ground truth



Feature	Pinned	Hovered
Size	1555	19174
race	Black	White
sex	Female	Male

Figure 1.14: After selecting different subgroups of interest, users can inspect the models' performance on these groups as well as their label balance in the ground truth. This can help in identifying possible bias causes of a model.

labels. Similarly, other tools suffer from the same limitations in their functionality for ground truth label analysis as they do in their functionality for prediction analysis. For instance, the WhatIf tool does not allow intersectional analysis based on more than two sensitive attributes (see section 1.4.1.2) and the RAI dashboard only gives visualizations (and no quantitative measures) on the ratio of positive/negative labels in the ground truth (see section 1.4.1.1). Extending the tools' functionality in these regards will thus be a way to accommodate for the design considerations addressed in Table 1.1 as well as Table 1.2.

Inspect relation between features *Historical bias* may not only be present in the ground truth labels but also the data itself [121]. Take for instance the *Adult dataset*, which may not only be unfair in terms of the unequal distributions of high/low income but also in terms of other features. To observe these inequalities, tools must enable auditors to inspect the relation between different features. Currently, this is already possible, using FairSight, the RAI dashboard and the WhatIf tool. Both in the RAI dashboard and the WhatIf tool, this functionality is provided by a visualization interface, where users can specify which variables should be plotted on the axes of a two-dimensional graph, along with other options for colouring or labelling data attributes according to the dataset's features (see Figure 1.10 (C) for the RAI dashboard and Figure 1.7 (A) for the WhatIf tool). Looking for instance at Figure 1.10 (C) we see interesting patterns in the education level of racial groups, and that people with an Asian, Asian American or Pacific Islander ancestry have higher education levels than black people in this dataset. As mentioned in section 1.4.1.3 this could be seen as a ground for explainable discrimination. In other words, the fact that black people have lower education levels explains their lower income and could consequentially justify a bank giving out fewer loans to them. However, as also mentioned in this section, an essential part of a bias audit is questioning where these inequalities in input data come from and whether they reveal patterns of historical bias, that lead to unfairness in an ML model [64, 128]. In our example, the differences in education levels could partly be due to differences in educational opportunities for people of different races, caused by unequal funding for schools and overall differences in access to resources (e.g., private tutors or high-quality books) [37]. As this reflects a larger pattern of systemic racism in the US, this can hardly be seen as fair and an auditor might wish to make up for this unfairness, by choosing bias-transforming measures (as mentioned in section 1.4.1.1) as their fairness-goal. Thus, having tools that supports the analysis of input data, are needed to identify historical bias as a cause of bias in an ML model, as well as to help auditors make well-informed choices about the fairness requirements of a model.

As mentioned before, FairSight also supports the (visual) analysis of the input data, which is done in two ways: first, per feature two histograms are provided to show how the distribution of this feature differs on the protected and unprotected group (in this case women vs. men, see Figure 1.4 (A) and 1.13a). This functionality can be useful to detect patterns like the one mentioned above, but in its current form works only for one binary sensitive attribute, which limits its usefulness. Additionally, FairSight provides a two-dimensional graph where a dimensionality-reduced version of the input data is visualized, and the protected and unprotected datapoints are colour-encoded (see Figure 1.5 (A)). While this can help in understanding how distinct both groups are overall, it is hard to understand where potential differences come from and whether they might indicate problematic historical biases.

Compare train and test set Another possible cause of bias in ADM systems is when the data a model is trained on, is not representative of the data it is applied on [120]. To give an example of this *representation bias* in the case of our loan application system: imagine a bank using relatively old data to train their ML model, where some population groups like non-male people are less represented than they are at the time of model deployment. Since the model does not see all population groups equally at training time, it will likely not perform accurately/fairly on the underrepresented groups once it is deployed. To establish representation bias as the cause of a model's unfairness, an auditing tool must let users compare the distribution of train- and test set. For this, tools must make a clear distinction between both so either can be individually inspected and then compared. Currently, none of our reviewed tools supports this functionality. All tools require the user to upload one dataset, along with its ground truth labels and the corresponding model's predictions. Users must choose whether this dataset is the same as the one the model has been trained on or is a separate test set. Since representation bias is a common cause of bias in ADM systems, this is a serious shortcoming in letting auditors identify this as a bias cause.

1.4.2.2 Inspecting subgroup sizes of interest

As we have touched upon in the previous section, a crucial part of the fairness audit of an ADM system, is ensuring that the audit is conducted on a representative test set [10,65,120]. After all, if a test set does not contain all the groups that an ADM system will be applied on, it is impossible to estimate whether the system will behave fairly on those groups. The way in which auditing toolkits can help in crafting representative test sets is by giving clear indications of the size of different subgroups in the data. We already see an implementation of this in FairVis, where in Figure 1.14 we see (along with some performance measures) the number of people represented in selected subgroups. Though this is already useful, it could be even more useful if the tool would allow users to order subgroups according to their size, just like it is already possible to order them according to performance/fairness measures. Also, in the WhatIf tool and the RAI dashboard it is possible to observe the group sizes of selected subgroups. However, in the WhatIf tool this is only possible for subgroups based on two sensitive attributes (see Figure 1.8 (b)) and in the RAI dashboard it is only possible to observe the subgroup size for one group at a time (see Figure 1.9 (c)). Functionality for an easier inspection of subgroup sizes would be useful.

1.4.3 Functionality to make bias detection scalable

The requirements introduced in section 1.4.1 and 1.4.2 relate to the tools' functionality to audit an ADM system's predictions and input data for bias. The requirements introduced in this section focus on making sure that this audit is scalable. On the one hand this refers to design consideration
DC_prioritize_systemic_biases [65,84,86], in that auditors do not want to inspect errors of an ADM system that are due to chance, rather than reflective of systemic bias issues. On the other hand, auditors fear that in focusing on only the big "obvious" biases, they might miss important blindspots; something that should be accounted for according to **DC_account_for_blindspots** [65,84].

Table 1.3: Requirements to make a bias audit scalable.

Tool makes bias detection scalable	
Tool let auditors narrow down the biases they need to inspect [DC_prioritize_systemic_biases]	
Report Confidence Intervals	/
Group similar subgroups together	DL RD
Group similar individuals together	FS
Tool let auditors narrow down the biases they need to inspect [DC_account_for_blindspots]	
(Sub)group biases	FV RD
Individual biases	FS

1.4.3.1 Tool let auditors narrow down the biases they need to inspect

The first way to make bias detection more scalable is by providing tools that can narrow down all the biases auditors need to inspect. In this section we explore how this can be accomplished.

Report confidence intervals An ADM model is unlikely to yield the same performance and the same positive decision ratio among all groups of interest. Hence, an important question in the audit of an ADM system is which disparities are due to chance and which ones reflect systemic issues. Hence, interviewees in the studies of [65] and [86] expressed their interest in tools that let them explore the statistical significance of subgroup biases, by reporting the confidence interval of bias measures. Currently, none of our reviewed tools supports this feature but some literature on subgroup fairness in ADM gives insight into how it can be implemented: Wang et al. apply the same model on five different test sets to calculate the confidence interval of the resulting bias metrics [130]. Similarly, Friedler et al. have studied how significant biases are, by calculating and comparing the bias metrics over multiple train-test-splits of a model [54]. Likewise, an interactive fairness tool could take a model, a train and a test set as input to then automatically divide the test set into different splits and calculate the confidence interval of the model's fairness measures over them. This would require more effort from the tool developers' side, since this tool needs to access more than just a simple CSV-file of the test data, but also the model itself. Alternatively, users themselves could provide the models' results on different test sets, over which a tool could (without needing access to the model) calculate the confidence interval. This approach would require more effort from the users' side in setting up a file containing all the necessary data. In choosing which option is more viable for an auditing toolkit, it is important to consider the current workflow of ADM model builders and auditors. Hence, before the feature of reporting confidence intervals is implemented in a tool, more conversations with practitioners would be needed to understand how they currently set up their model evaluation and how an auditing toolkit could account for that.

Group similar subgroups together Though not directly suggested by possible auditors, but already implemented in some tools, another way to reduce the number of subgroups an auditor needs to inspect, is to (automatically or manually) group similar subgroups together. The RAI dashboard already allows one to do so: instead of e.g. generating one subgroup of 50-year-old men, and another of 51-year-old men, users can generate a subgroup of men with a certain age range (e.g. 50-55) and inspect a model's performance on it. Similarly, when working with categorical features, users can group people with similar feature values. In the *Adult dataset*, there is for instance a variable "workclass" with values like "Federal Government" "State Government" or "Local Government". Instead of inspecting each subgroup individually, users of the RAI dashboard can generate one subgroup of all people working for the government (see Figure 1.9 (b)) for the dashboard's subgroup generation component). Similar subgroups are also automatically grouped together in DiscriLens: in Figure 1.3 (e), the group of people with more than 65 workinghours per week is suggested as a discriminatory itemset, allowing an auditor to inspect a bigger group of people than when only looking at the group of people with exactly 65 workinghours per week. Grouping similar subgroups together is currently not possible in the other tools supporting the analysis of subgroup biases (i.e. FairVis and the WhatIf tool). Hence, implementing this feature could help in making the bias analysis more scalable.

Group similar individuals together When it comes to individual biases in an ML system, it is even more unfeasible for an auditor to inspect all of them, since individual biases only affect one data instance at a time. A way in which tools can help is to group similar instances facing discrimination together. FairSight is the only tool that currently does so, in its Feature Inspector tab (see Figure 1.6). Here, each feature is visualized in a histogram, and data points that face high levels of individual discrimination are marked in red, to observe their value for the given feature. To illustrate, in the histogram showing the data distribution on the feature "sex", we see that on this feature most individually discriminated instances have the value "female". Auditors can use this histogram to study patterns of individual discrimination in a quick and scalable way.

Another way in which the inspection of individual biases could be made more scalable is by applying a clustering algorithm on instances that score high on an individual discrimination score. The auditor could inspect the resulting clusters to find common patterns of individual bias. A similar approach was once adopted by Luong et al., who first identified individually discriminated instances, and then derived decision rules to learn what sets them apart from other instances [89]. Linking individual examples of discrimination to more general "discrimination rules", can also be a promising way to facilitate auditors' understanding of global discrimination patterns in the model [109]. Hence, this kind of functionality could be added to tools that allow the inspection of individual discrimination (i.e. RAI dashboard and the WhatIf tool), but where this inspection needs to be done "one at a time".

1.4.3.2 Tool automatically highlights most important biases

Auditors are concerned that in efforts to make a bias audit scalable, they will miss hidden but important patterns of bias (**DC_account_for_blindspots**) [65, 84]. Generally, they know that by involving stakeholders in the auditing process, as well as having diverse development teams they are more likely to identify the different population groups that may be subject to bias [65]. Still, they also find it useful when a tool can automatically suggest patterns of bias that would otherwise go unnoticed. In this section, we discuss ways in which both group and individual biases can be automatically highlighted by our toolkits.

Group Biases Currently, only FairVis and the RAI dashboard automatically suggest subgroups that may be affected by bias (see Figure 1.2 (E) and Figure 1.15, for a close-up of this feature). We see that the user has previously generated subgroups based on combinations of people's sex and race, and we already see the false negative rates for these groups, including, e.g., the groups of black women and white women. In the "suggested subgroups" tab we see other potential subgroups of interest, sorted according to their False Negative Rates. In this case, the group of divorced black women from the United States, as well as the group of unmarried white women working in the private sector are suggested. When clicking on these suggestions, we indeed see that they have considerably high False Negative Rates, also compared to their supergroups of black women and white women.

Similar suggestions are also given in the RAI dashboard, as part of the "Error Analysis" component (Figure 1.9 (C). In this case, the user has selected to specifically look at the subgroup of white men. Within this group, they want to inspect for which subgroups the recall is especially high or low. On top of the error tree, we can see that the overall recall for the group of white men is 0.29. When following one of the paths of the tree we see that this score is much lower (0.0) for the group of white men older than 31 and with an education level below 7 (where an education level of 7 describes people who have followed education until 11th grade of high school).

The suggestions of FairVis and the RAI dashboard can help auditors in detecting otherwise missed patterns of subgroup bias, but also contextualizing these biases. Thus either of these features can be a useful addition to other auditing toolkits. Still, it should be noted that their current implementation of subgroup bias suggestion may not be ideal.

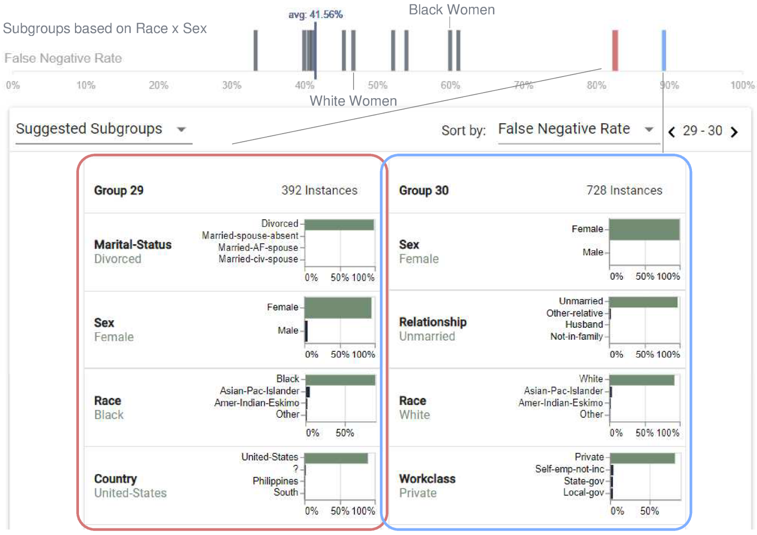


Figure 1.15: FairVis is one of the tools that automatically suggest potentially discriminated subgroups to its users. In this case it suggests the group of divorced black women from the US as well as the group of unmarried white women working in the private sector as two groups with high False Negative Rates.

First, both tools only show the performance measure, without giving absolute measures about how many people are affected by unfair treatment. Consider, for instance, the group of divorced black women from the US, that is suggested by FairVis. Inspecting the corresponding data and the ADM model's predictions on it, we found that out of a total of 32k dataset instances this subgroup consists of 300 people. Out of these, only 22 have a positive label (i.e. high income) in the data. Since the model only correctly predicts this label for 3 people, the False Negative Rate is so high for this group. While this is certainly a problematic pattern, it only concerns a very small part of the data, which the tool does not make apparent. It is then also hard to estimate whether this bias is significant for the group of black divorced women from the US, or whether it is a pattern coming forth from the general bias against black women (independent of their marital status and nationality).

The second concern about how discriminated subgroups are automatically suggested in the RAI dashboard and FairVis, is that both tools only suggest subgroups based on error-based bias measures. To improve this functionality, it is important that other forms of bias, e.g. defined by outcome-based measures (see section 1.4.1.1), are also automatically highlighted.

Individual Biases As mentioned in 1.4.1.1, the only tools that support the detection of individual bias are FairSight, the WhatIf Tool, and the RAI dashboard. Out of all, only FairSight automatically highlights some cases of individual bias. This is done through the distortion matrix (see Figure 1.5 (D)). To reiterate, the distortion between two pairs of individuals is high when they are close on the input space (i.e. they are similar in terms of their features), but distinct on the output space. Instance pairs with high distortion are highlighted with different colour saturation than instance pairs with low distortion. Additionally, instance pairs that differ on the sensitive attribute "sex", are coloured differently than instance pairs with the same sex. Hence, to find potentially discriminated instances, users could look for female-male instance pairs with high distortion (as their difference in output might only be explained by sexist biases, as the instances are otherwise close in input space).

No clear cases of individual discrimination are highlighted in the WhatIf tool and the RAI dashboard. One possibility to add this functionality is by using the tool's "what if" analysis and automatically highlighting cases that experience a change in their prediction outcome if the value of a sensitive attribute or a redlining attribute is changed.

1.5 Conclusion & Future Research

In this paper, we have presented an overview of the functional requirements that users have for auditing toolkits. We evaluated six available toolkits according to these requirements and identified realistic ways to overcome their shortcomings. One of the most common shortcomings is their lack of flexibility: many tools assume that there is only one binary sensitive attribute that auditors need to assess for possible biases, and that information on this attribute is always available. To address this issue, tools must be developed that make less rigid assumptions about the availability and number of sensitive attributes. Such toolkits allow for audits where discriminatory bias occurs solely on the basis of proxy attributes and where bias may be of intersectional nature. Other important design requirements include the tools' functionality to assess the training data for possible bias causes and the extent to which they make audits scalable.

While our requirement checklist and the tool evaluation can already guide developers in creating better and more suitable tools, some aspects still should be studied to unlock their full potential.

Integration in workflow The tools that we reviewed differ in the way they need to be set up. Some are webtools that take CSV files of the data and models' predictions as input [22, 115] others are evoked through python libraries [116,135]. Whether practitioners choose to use toolkits in practice will depend on the ease with which they can be integrated into their workflow. The ADM developers that were interviewed by Lee & Singh, for instance, preferred tools that can be evoked in Python and that integrate well with other python libraries like pandas or sklearn [86]. Additionally, they had privacy concerns about using web-based tools, that require them to upload sensitive data on external sites [86]. The preferred way of setting up tools likely also varies, depending on the technical skills of an auditor and whether they are involved in system development or not [90]. Hence, developers should spend serious effort on designing tools that easily integrate into different types of workflows and that can run on local machines to minimize privacy concerns.

The usability of the tools Auditing toolkits should offer the right functionality, but at the same time, this functionality should be easy to use. To determine the current

Sex Categories				
	# of Applicants	# of Selected	Selection Rate	Impact Ratio
Male	1390	667	48%	1.00
Female	1181	555	47%	0.979

Table 1.4: An example of the type of report that is mandated by the Local Law 144.

usability of tools, more studies are needed. Some tool developers, already conducted (small) usability studies as part of their research papers [3,133,135]. However, all these studies suffer from some disadvantages like only testing the tools on university students with Computer Science backgrounds and only giving the participants tasks that are specific to the exact purpose of each tool. To illustrate, one of the tasks of the usability study of FairSight was *"Can you quantify the degree of fairness in the ranking outcome?"* [3], which requires participants to locate one specific fairness metric reported in the tool. While it makes sense to study the usability of specific tool's components, it is worthwhile to give users more general tasks, that provide a better understanding of how each tool would be used "in the wild". Inspiration for task set-ups could be taken from literature on the interpretability of explainable AI (xAI). For instance, Kaur et al. purposefully manipulated the predictions of a Machine Learning model and studied how well explanation methods could help participants in identifying these undesirable patterns. A similar approach could be taken for testing the usability of bias auditing tools, to see whether the tools help in finding unfair patterns in a model (as well as contextualizing these patterns and identifying their causes) [76].

Need for clear legislation We have already mentioned how auditing toolkits are highly relevant in light of upcoming legislation like Local Law 144 and the EU AI Act. To shortly summarize: the EU AI Act calls for human oversight of decision-making systems to evaluate risks concerning health, safety, and fundamental rights. Though non-discrimination is considered a fundamental right, the act does not provide specific guidelines on how to ensure that a system does not violate it [46]. New York's Local Law 144 is a bit more precise in the requirements it imposes: it mandates that any algorithm used for hiring decisions must undergo an audit by an independent third party, with the results made publicly available. The audit should assess the algorithm's output for potential discrimination based on sex, race and their intersection and must report some basic measurements regarding the corresponding population groups. To specify, it needs to report the number of job applicants for each group, their selection rate (i.e., the percentage of applicants that are selected to move forward in the hiring process) and their impact ratio. The impact ratio is calculated as the selection rate for the group, divided by the selection rate of the highest selected group. An example of such a report as mandated by the law is given in Table, where the different measures among sexes are given 1.4 [8].

While Local Law 144 is more specific than the EU AI Act, it still lacks (similarly to the EU AI act) specific standards for how the success or failure of an audit is determined. Additionally, neither of the new legislation gives further information on how possible signs of discrimination should be further inspected, e.g. by contextualizing differences in selection rates or by trying to find their cause in the training data. On the one hand, not having clear definitions of when an audit passes or fails helps in accommodating many use cases, where the most sensible bias measure differs depending on the decision task (see section 1.4.1). On the other hand, researchers and auditors have warned that the lack of more elaborate standards creates a risk of companies conducting minimal, superficial audits, that merely fulfil the regulatory requirement without genuinely addressing bias issues [35]. As hopefully, more rigid legislation will arise, the requirements in auditing toolkits will evolve accordingly. Until then, we also believe that the currently available toolkits and the requirements auditors have in them can shape the new rules that should be set into place. Some of the requirements highlighted in our paper, such as tools' functionality to inspect the training data for different forms of bias and the ability to contextualize biases, are related to essential components of the overall auditing process. Hence, our study and other relevant research contributions can serve as a basis for determining the best practices for audits, which could in turn be incorporated into new laws.

Chapter 2

Measuring Individual Fairness

In this chapter, we deal with one important part of bias audits, namely the measurement of individual fairness, which can be done through situation testing. This method originates from the social sciences, where it is used to assess discriminatory practices by comparing the treatment of individuals in similar scenarios where only one variable of interest (such as race, gender, or age) differs. In the data-driven equivalent of this practice, the goal is to identify similar instances in the dataset, that exhibit significant differences in their historical decision labels or prediction labels of an ADM model. A crucial and non-trivial component of this approach is defining a suitable distance function to determine similarity. This distance function must disregard attributes irrelevant to the decision problem and weigh other relevant attributes appropriately. In this chapter, we show how such a distance function, in the form of Weighted Euclidean distance, can be automatically learned from the data without relying on external resources like Causal Bayesian Networks or lengthy human annotation processes. We demonstrate how this new way of defining distances improves the performance of current situation testing algorithms, especially in the presence of irrelevant attributes.[1]

2.1 Introduction

As we have emphasized both in the Introduction and Chapter 1 of this thesis, one crucial component of any bias audit lies in the measurement of individual fairness, both in the historical decision labels of the training data and an ADM model's predictions. Though group fairness measures like *demographic parity* or *equal opportunity* are important to understand global patterns of discrimination, individual fairness measures give a more nuanced view of whether differences in treatment between demographics may be justifiable by other features [128]. Further, enforcing individual fairness can prevent

[1] This chapter is based on: *Lenders, D., & Calders, T. (2021, September). Learning a fair distance function for situation testing. In Machine Learning and Principles and Practice of Knowledge Discovery in Databases. ECML PKDD 2021. (pp. 631-646).*

cherry-picking, i.e. blindly distributing positive decision outcomes in the pursuit of some group fairness goal without paying further attention to whether these positive decisions make sense on an individual level [42]. Lastly, as we have also found in our literature review in Chapter 1, auditors find it easier to understand global patterns of unfairness, when being given examples of how this discrimination looks on an individual level [41, 109] (see Chapter 1.4.1.1). The goal of this chapter is, therefore, to zoom into individual fairness and show how it can be measured in a context-dependent way, appropriate to a given decision task.

In social sciences, individual fairness has often been measured through “situation testing", where two nearly identical individuals, that only differ on one sensitive attribute (like their gender or ethnicity), are put in similar situations, like applying for a loan, and their difference in treatment is observed. This principle was translated into an algorithm by Luong et al. [89]. To illustrate their methodology, let us revisit a toy dataset for a loan allocation setting, with “race" being the sensitive attribute and “black people" being the historically non-privileged group.

Table 2.1: Dataset of an illustrating example.

#	Race of Owner	Zip code	Credit History	Credit Amount	Loan Approved?
1	Black	1234AZ	Had Default	20-30k	No
2	Black	1234AZ	No Defaults	20-30k	Yes
3	White	4567BY	No Defaults	20-30k	Yes
4	White	4567BY	Had Defaults	10-20k	Yes

Say we want to find out whether instance 1 was decided not to be granted a loan, because of discriminatory bias against their race. Following Luong et al., we find its most similar dataset instances, k from the non-privileged (black people) and k from the privileged group (white people), and compare the positive decision ratio of both. If this ratio is significantly lower for the nearest black than nearest white neighbours, the instance is flagged as potentially discriminated. While the idea behind this algorithm is appealing, it is challenging to accurately define distances in the data. First, Luong et al. state that distances should only be defined on attributes relevant to the decision problem, but it is not trivial which attributes can be seen as such. Second, in their distance function, all attributes contribute equally to the defined distances, when it is desirable that features more relevant to the problem also have a higher weight in the function.

Recognizing these shortcomings, Zhang et al. [144] recently proposed a more refined approach for situation testing in which the distance function takes the causal relationships between the features and the decision attribute into account: features that have a higher causal effect on the decision attribute will also contribute more to the distances between dataset instances. If, for instance, in the above dataset, a good “credit history" is the most crucial factor for being eligible for a loan, this feature should also have the highest weight in the distance function. Though in theory, this approach works well, in practice it is hard to find an appropriate causal network to use for this method: networks defined by experts are not always available, while network learning algorithms do not always yield accurate or robust results [38].

Next to Zhang’s work, other studies discuss ways to learn fair distance measures, basing this process on human feedback [69, 96, 131]. While incorporating human expertise can be appealing, this methodology can be time-consuming and prone to biases. For example, Wang et al. [131] propose learning a distance function based on human ratings of the likelihood that 200 data instances should receive a positive decision outcome. The goal of this annotation process is to find which features humans find important for a decision task, such that these can be appropriately incorporated into the distance function. Alternatively, Ilvento [69] suggests learning a distance function by querying human opinions on the similarity of multiple pairs of instances. As neither of these annotation tasks is very trivial, and annotators may also be inconsistent in their ratings it is hard to guarantee that a suitable distance function is learned.

To learn a task-appropriate distance function without external resources, this chapter deals with how a function can be directly learned from the data. Rather than relying on a causal network or human input, these distances are learned through an optimization algorithm, which learns the parameters for the Weighted Euclidean distance, such that features with the highest impact on the decision label contribute most to the distance. At the same time, the distance function is "fair" in the sense that it does not discriminate between members of the privileged group and the unprivileged group. That is, if members of both groups differ on attributes irrelevant to a decision problem, these attributes will not contribute to the distance between the two instances. We demonstrate the superior performance of our proposed distance function on both simulated and real-life data. In addition to defining a new distance measure, we refine Luong's situation testing algorithm, by proposing methods for selecting its hyperparameters. While the main purpose of the resulting algorithm is to detect individual discrimination, it could also be applied for discrimination prevention. To do so, one could use the algorithm to first detect and then remove discrimination in the data, such that classification algorithms trained on it will be fair from the start.

In this chapter, we will not focus on this task, but will solely concentrate on situation testing for discriminating detection. Futher, we emphasize from the start that this method is not meant to give legally binding judgements about whether an individual is discriminated against, but is meant as a tool to flag potentially discriminated instances. Whether the similarity between an instance and its neighbours is sufficient to base discrimination decisions on, should still be determined by a human auditor.

2.2 Analysis of Existing Situation Testing Methods

Before we go into detail about the existing methods of situation testing, let us first reintroduce some of the notations of our introduction, that we will use throughout the rest of this chapter as well:

- X is a dataset, consisting of N individuals. We use the notation $\mathbf{x}$ to refer to one individual of the dataset
- A is sensitive variable (e.g. gender, race, religion, ..), which for this chapter we assume to be binary. We use $S(\mathbf{x}) = +$ to denote that $\mathbf{x}$ belongs to the demographic group we consider to be privileged by society (e.g., men) and $S(\mathbf{x}) = -$ to denote that $\mathbf{x}$ belongs to a non-privileged group (e.g., women)

- D is a decision outcome associated with each individual, where we use $D = 1$ and $D = 0$ respectively, to denote the favourable and non-favourable decision outcome
- $\mathcal{G}$ are the attributes of the dataset that are legally grounded for being used in the decision-making process. It is assumed that these attributes are given beforehand. We use $G(\mathbf{x})$ to denote the value of $\mathbf{x}$ for attribute $G \in \mathcal{G}$

2.2.1 Situation Testing - Luong et al.

The basis of Luong's situation testing algorithm is, to deem a protected instance with negative decision label as discriminated, if there is a considerable difference in the ratio of positive decision labels among its non-privileged and privileged neighbors. The similarity between two people x and y is defined as the sum of value differences between all legally grounded attributes of x and y, where VD_G is an appropriate distance for the domain of G.

$$d(\mathbf{x}, \mathbf{y}) = \sum_{G \in \mathcal{G}} VD_G(G(\mathbf{x}), G(\mathbf{y})) \tag{2.1}$$

Luong et al. propose to use this distance function with a kNN classifier to define discrimination scores. Each protected individual with a positive decision label, cannot be discriminated and thus gets a discrimination score of 0. For each protected individual with negative decision label, we define the discrimination score as the difference in the ratio of positive decision labels among its k nearest privileged and non-privileged neighbors. If this score is higher than threshold t, we decide that the instance in question was discriminated.

Shortcomings Luong's approach Luong's approach elegantly simulates the method of situation testing. However, in some circumstances, their defined distance function may not capture the full complexity of the problem. Consider e.g., the dataset displayed in Table 2.1. In a fair setting, only applicants who pose a large risk of defaulting, should not be considered eligible for a loan. The first question is which attributes of the data to include in $\mathcal{G}$, i.e. the set of legally grounded attributes used in the distance function. While it is clear that "Race" should not be considered, the feature "Zip code" provides more ground for discussion: on the one hand, the Zip code of a property might be a proxy for someones' financial means, which is fair to consider when deciding on loan eligibility. On the other hand, it can be seen as a "red-lining attribute", i.e. an indicator of "Race" that can be used for discriminatory practices and should therefore not be included in $\mathcal{G}$.

Even when excluding "Zip code" from the legally grounded attributes, Luong's approach may be too simplistic to define similarity. For instance, we could ask ourselves whether dataset instance 3 or 4 is more similar to instance 1. According to Luong's distance function, both pairs are equally similar, because instance 1 and 3 only differ on their credit history and 1 and 4 only on their credit amount. In reality, we however want to put different emphasis on both features, depending on how they relate to a persons' loan eligibility.

2.2.2 Situation Testing - Zhang et al.

Recognizing the above shortcomings, Zhang et al. recently proposed a more refined method for computational situation testing, where the distance function is partly based on causal relationships found in the dataset [144]. Given a Causal Bayesian Network (*CBN*), that models these relationships, Zhang et al. propose to define the distance between two dataset instances solely based on attributes that have a direct causal effect on the decision attribute. Again this distance is measured as the sum of value differences for each attribute, but this time these value differences are weighted based on their causal relationship with the decision attribute. This causal relationship is measured by the "intervention", in which an instance's value on the attribute of interest is replaced by the other instance's value for that attribute. By calculating the difference in the probability of a positive decision label before and after this intervention, we get an idea of how the attribute change causally affected this probability. For the exact formulas of Zhang's approach, we refer to their paper [144].

Using this refined distance function, we again select the k nearest protected unprotected neighbors of a possibly discriminated instance. Again a discrimination score of an instance is defined as the difference in the ratio of positive decision labels between both groups.

Shortcomings Zhang's approach While Zhang's approach elegantly solves some of the problems of Luong's method, it suffers from the disadvantage that it heavily relies on a causal network to define the distances between dataset instances. Causal networks given by domain experts may not always be available or accurate, while networks learned by algorithms may not be very robust [38]. Since the distance function of Zhang et al. is only defined on the attributes that have a direct causal effect on the decision attribute, the presence or absence of a causal link can make a tremendous difference on the distances defined with this approach. Thus the results of Zhang's algorithm may vary considerably, depending on the causal network they were based on.

2.3 Learning a Fair Distance Measure

To overcome the shortcomings of Luong's and Zhang's distance measures, we propose a way in which a distance function can be learned from the dataset. This function defines distances of instances on the value differences between their features, as well as on the importance of these features for the decision label. We ensure that features only correlated to the decision attribute through a sensitive attribute, do not contribute to any distances. To guarantee robust results, we do not rely on a causal network to define the distance function but rather use an optimization algorithm to learn it. The distance metric we are going to use for this task is the Weighted Euclidean distance, given by equation (2.2). Note that to include nominal variables in the distance function these variables first need to be one-hot-encoded. We assume that next to the label D and the sensitive attribute S, there are numerical attributes $B_1, \ldots, B_n$. In order not to overload notation we will denote $B_i(\mathbf{x})$ by x_i. We define the weighted Euclidean distance in the usual way; for a vector of weights $\mathbf{w} = (w_1, \ldots, w_n)$, the distance is:

$$d_{\mathbf{w}}(\mathbf{x}, \mathbf{y}) := \sqrt{\sum_{i=1}^{n} w_i (x_i - y_i)^2} \tag{2.2}$$

In this function, the distance between two instances is defined as the sum of squared differences between its features. These differences are multiplied by a weight, which indicates how important the corresponding feature is for the overall distance. The question that now arises is how to learn these weights. The basic idea behind our optimization algorithm is that they should be learned such that the distances between instances with the same class label are small while they are big between instances with different class labels.

To make sure that the sensitive attribute does not directly contribute to the distance between two instances, the distance function $d_{\mathbf{w}}$ will only be defined over the non-sensitive features of the dataset. However, this alone does not make sure that the distance function does not differentiate between the protected and unprotected group through "red-lining attributes". These are attributes like "height" which are strongly correlated to "gender" and therefore (in case of discrimination) also correlated to the decision attribute. Despite this correlation, red-lining attributes should not contribute to the distance between two instances, and should therefore be assigned low weights in the learning process. An easy way to ensure this is splitting the dataset X according to sensitive attribute S when minimizing the distance between instances with the same class label, and maximizing the distance between instances with a different class label. With this approach, our distance function is given by $d_{\mathbf{w}^*}$ where $\mathbf{w}^*$ is the vector of weights that minimizes the objective function given in (2.3). We here use S^- and S^+ to refer to the instances belonging to the non-privileged and privileged group respectively. Further, we use $|S|$ to denote the number of dataset pairs with the same class label and $|U|$ to denote the number of dataset pairs with a different class label.

$$\begin{aligned} \Bigg\{ &\frac{1}{|S(A^-)|} \sum_{\substack{\mathbf{x},\mathbf{y} \in A^- \\ D(\mathbf{x})=D(\mathbf{y})}} d^2_{\mathbf{w}}(\mathbf{x}, \mathbf{y}) + \frac{1}{|S(A^+)|} \sum_{\substack{\mathbf{x},\mathbf{y} \in A^+ \\ D(\mathbf{x})=D(\mathbf{y})}} d^2_{\mathbf{w}}(\mathbf{x}, \mathbf{y}) - \\ &\frac{1}{|U(A^-)|} \sum_{\substack{\mathbf{x},\mathbf{y} \in A^- \\ D(\mathbf{x})\neq D(\mathbf{y})}} d^2_{\mathbf{w}}(\mathbf{x}, \mathbf{y}) - \frac{1}{|U(A^+)|} \sum_{\substack{\mathbf{x},\mathbf{y} \in A^+ \\ D(\mathbf{x})\neq D(\mathbf{y})}} d^2_{\mathbf{w}}(\mathbf{x}, \mathbf{y}) + \lambda \|\mathbf{w}\|_2 \Bigg\} \end{aligned} \tag{2.3}$$

In this function the term $\lambda \|\mathbf{w}\|_2$ acts as an L2 regularizer, which forces weights not relevant for the task to be close to zero. It is necessary to include this regularizer because otherwise the weights of irrelevant attributes would neither increase nor decrease the value of the objective function, since they are likely to be equally distributed among instances with positive and negative class labels.

2.3.1 Theoretical Justification of the Distance Optimization Problem

In this part we show that the theoretical optimal solution for the optimization problem we defined in Equation 2.3 has several desirable properties, showing that it resolves some of the problems we identified in the approach of Luong et al., without relying on a *CBN* to do so.

The objective function is an estimator of the following risk function together with a regularization term $\rho(W) := \lambda \|\mathbf{w}\|_2$:

$$R(\mathbf{w}) := \sum_{a=0,1} \Big(E[d^2_{\mathbf{w}}(\mathbf{x},\mathbf{y}) \mid D(\mathbf{x}) = D(\mathbf{y}), A(\mathbf{x}) = A(\mathbf{y}) = a] - E[d^2_{\mathbf{w}}(\mathbf{x},\mathbf{y}) \mid D(\mathbf{x}) \neq D(\mathbf{y}), A(\mathbf{x}) = A(\mathbf{y}) = a] \Big)$$

All expected values are taken over the distribution that generated the dataset.

The next result shows that if an attribute only contributes to the label through the sensitive attribute, then its weight will be 0 in the optimal solution. An attribute B "only contributing through the sensitive attribute" means that the label D is conditionally independent from B given the sensitive attribute.

Theorem 1. *Let B be an attribute such that $D \perp B|A$, and let $\mathbf{w}^*$ be such that $R(w^*) + \rho(w^*)$ is minimized. Then $B(\mathbf{w}^*) = 0$.*

Proof. Assume for the sake of contradiction that $B(\mathbf{w}^*) \neq 0$ and let $\mathbf{w}'$ be the vector with $C(\mathbf{w}') = C(\mathbf{w}^*)$ for all attributes $C \neq B$ and $B(\mathbf{w}') = 0$; that is: we get $\mathbf{w}'$ by setting the weight corresponding to B to 0.

Clearly, $\rho(\mathbf{w}') < \rho(\mathbf{w}^*)$; $\rho(\mathbf{w}^*) = \rho(\mathbf{w}') + |B(\mathbf{w}^*)|$. We will show now that $R(\mathbf{w}') = R(\mathbf{w}^*)$. This is easy to see; first observe that:

$$E[d^2_{\mathbf{w}*}(\mathbf{x},\mathbf{y})] - E[d^2_{\mathbf{w}'}(\mathbf{x},\mathbf{y})] \quad = \quad E[B(\mathbf{w}^*)(B(\mathbf{x}) - B(\mathbf{y}))^2]$$

Then, since B is independent of D conditioned on A,

$$\begin{aligned} & E[B(\mathbf{w}^*)(B(\mathbf{x}) - B(\mathbf{y}))^2 | D(\mathbf{x}) = D(\mathbf{y}), A(\mathbf{x}) = A(\mathbf{y}) = a] \\ = \; & E[B(\mathbf{w}^*)(B(\mathbf{x}) - B(\mathbf{y}))^2 | A(\mathbf{x}) = A(\mathbf{y}) = a] \\ = \; & E[B(\mathbf{w}^*)(B(\mathbf{x}) - B(\mathbf{y}))^2 | D(\mathbf{x}) \neq D(\mathbf{y}), A(\mathbf{x}) = A(\mathbf{y}) = a] \end{aligned}$$

Combining these pieces we get:

$$\begin{aligned} R(\mathbf{w}^*) - R(\mathbf{w}') &= \sum_{a=0,1} E[d^2_{\mathbf{w}^*}(\mathbf{x},\mathbf{y}) - {d'_{\mathbf{w}}}^2(\mathbf{x},\mathbf{y}) \mid D(\mathbf{x}) = D(\mathbf{y}), A(\mathbf{x}) = A(\mathbf{y}) = a] \\ &\quad -E[d^2_{\mathbf{w}^*}(\mathbf{x},\mathbf{y}) - {d'_{\mathbf{w}}}^2(\mathbf{x},\mathbf{y}) \mid D(\mathbf{x}) \neq D(\mathbf{y}), A(\mathbf{x}) = A(\mathbf{y}) = a] \\ &= \sum_{a=0,1} E[B(\mathbf{w}^*)(B(\mathbf{x}) - B(\mathbf{y}))^2 \mid D(\mathbf{x}) = D(\mathbf{y}), A(\mathbf{x}) = A(\mathbf{y}) = a] \\ &\quad -E[B(\mathbf{w}^*)(B(\mathbf{x}) - B(\mathbf{y}))^2 \mid D(\mathbf{x}) \neq D(\mathbf{y}), A(\mathbf{x}) = A(\mathbf{y}) = a] \\ &= \sum_{a=0,1} E[B(\mathbf{w}^*)(B(\mathbf{x}) - B(\mathbf{y}))^2 \mid A(\mathbf{x}) = A(\mathbf{y}) = a] \\ &\quad -E[B(\mathbf{w}^*)(B(\mathbf{x}) - B(\mathbf{y}))^2 \mid A(\mathbf{x}) = A(\mathbf{y}) = a] \;=\; 0 \end{aligned}$$

Given this, we know that $R(\mathbf{w}^*) + \rho(\mathbf{w}^*) > R(\mathbf{w}') + \rho(\mathbf{w}')$, which contradicts the optimality of $\mathbf{w}^*$. Hence, in an optimal solution $B(\mathbf{w}^*) = 0$ has to hold. □ □

2.4 Learning the Distances and Setting Hyperparameters

In this section, we will show which algorithm was used to learn the Weighted Euclidean distance. After learning, the function could be applied in the original situation testing algorithm. Here we will, however, introduce additional adjustments to the algorithm, which should help improve its performance.

2.4.1 Learning the distance function

To learn the Weighted Euclidean distance, the objective function given by (2.3) had to be optimized. To do so we applied the "SLSQP" algorithm, implemented in Python's `SciPy` library. This is a quasi-Newton method, that assumes that the region around the optimum of the objective can be approximated by a quadratic function. The first and second derivatives of the objective are used to find the stationary point of this function.

2.4.2 Selecting neighbors from the privileged group only

Luong's and Zhang's discrimination scores are calculated as the difference in positive decision labels between an instance's nearest privileged and nearest unprivileged neighbors. While this approach is not necessarily wrong, it moves away from the original idea behind situation testing, where we only observe how a member of a non-privileged group is treated differently than similar members from the privileged group. We argue that there is no need to look at more than one non-privileged instance at a time since we are only interested in how their decision label is different from privileged counterparts, not how it relates to decision labels of other non-privileged instances. After all, the motivation behind individual measures of fairness is that an individual can be discriminated based on their sensitive attribute even if there occurs no discrimination on a group level [42]. Based on these arguments, we suggest adapting the situation testing algorithm by deriving an instance's discrimination score only on the positive decision rate among its k privileged neighbors. We will later on refer to this as the "Situation Testing k" approach as opposed to the original "Situation Testing k+k" approach. Note, that a downside of this new approach is that high discrimination scores may be given too easily: imagine for instance a group of equally capable males and females of which 80% received a job offer, regardless of their gender. In this case, the 20% of women receiving the negative label will be wrongly flagged as potentially discriminated. We will reduce this potential downside, by providing a method of choosing a threshold (that turns discrimination scores into discrimination labels), that counteracts this effect (see section 2.4.4). Further, we will see in the experiment section how the adapted approach has positive effects on the performance of the algorithm and how it also helps in selecting the hyperparameters of the algorithm.

2.4.3 Setting the number of selected neighbours

One issue that previous works only slightly touch upon is how to choose a good value for k. Intuitively, k should be big enough to guarantee that the class information we gather from an instance's neighbors is representative. At the same time, k should not be too big, otherwise, the selected neighbors may not be close to the instance in question anymore. In this section, we try to quantify this intuition. For our reasoning, we make use of our choice of the previous section, where we propose to infer the non-biased decision label of a non-privileged group member from the decision labels of its privileged neighbors. Since there is no way to say whether a derived non-biased class label is correct, we could look at privileged group members instead. Since we assume that no discrimination occurs in this group, any class label that is derived for a member of this group should be the same as their actual class label in the dataset. In other words, we use a regular kNN classifier (that utilizes the desired distance function) to predict the labels of any privileged instance from the rest of the privileged group and see which k yields the highest accuracy in this approach. For our experiments, we choose to set the possible k-values to $\{10, 20, 30, 40\}$. This approach can simultaneously be used to choose the best value for λ for the learned distance function, by checking which combination of k and λ works best for an accurate prediction of the decision labels in the privileged group.

2.4.4 Setting the threshold

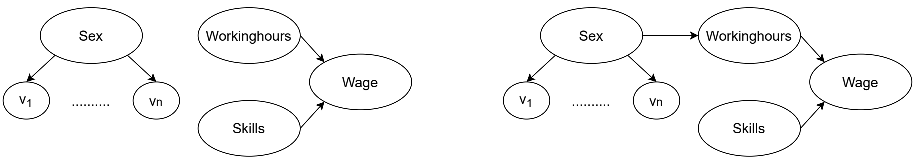
In the situation testing algorithm, we turn discrimination scores into discrimination labels by checking whether they exceed a given threshold t. It is proposed to either base t on existing discrimination laws, or to let the analyst choose an appropriate value for it. However, statistic-based discrimination laws do not always exist, and so far there is no clear guideline on how to adopt a general approach for choosing t. The idea behind our alternative approach for selecting t is that any difference in how a non-privileged member was treated differently than its privileged neighbours, can only be interpreted as discrimination if this difference is higher than expected by chance. To quantify this idea we could look at the discrimination scores that are assigned to the privileged members of the dataset with a negative decision label. Since we assume that they were not discriminated, any discrimination score higher than 0 reflects some randomness in the data or the general chance of receiving a positive decision label. To make sure that the discrimination labels assigned to the non-privileged individuals reflect more than bare randomness, we could thus choose a threshold that lies higher than most discrimination scores assigned to the privileged indices. This threshold could for instance be at the maximum non-outlier value of the privileged discrimination scores. This value can easily be found by making a boxplot of discrimination scores of the privileged members with negative class labels and setting the discrimination-label threshold at its upper whisker.

2.5 Experiments on Simulated Data

2.5.1 Generation Process of Data

Our first experiments were conducted on simulated datasets, based on the causal networks displayed in Figure 2.1a and 2.1b. In these datasets the “Wage" of an employee is the decision attribute, which either can be “high" or “low". “Sex" is taken as the sensitive attribute, where women represent the non-privileged group and men the privileged group. Further, the skill level and amount of working hours of each employee are recorded, where a higher value for either of them increases an employee’s chance of receiving a high wage. Since in the dataset based on Figure 2.1b, there is a link between “Sex" and “Workinghours" (with women being likely to work fewer hours) this dataset contains what we call “explainable discrimination": even though men receive on average a higher wage than women, this difference is not seen as a form of illegal discrimination, since it can be explained by other variables of the data. Note, that this is an oversimplified example used for illustration purposes only. In real life these relations are much more complex, and a presence of an “explanatory attribute" may not fully justify a different treatment between members of a privileged and an unprivileged group.

In the datasets based on each network, we added 1, 4 or 7 features that are only correlated to the persons’ sex. In our experiments, we assume that these attributes belong to the set of legally grounded ones (and will be used in Luong’s distance function), even though they do not give information about the decision label. The formulas used to generate the data are given in Figure 2.2. Starting from the non-biased datasets, we simulate a situation where a decision-maker is biased against women. This person decides that the wage of several women should be “low", when in fact it was supposed to be “high". To



(a) Non explainable discrimination (b) Explainable discrimination

Figure 2.1: Causal networks used for data simulation to generate the unbiased data. This data is used as the ground truth in the experiments. A version of this ground truth in which discrimination is added is used for testing the algorithms.

Figure 2.2: The formulas to simulate the data corresponding to CBNs in Figure 2.1a and 2.1b. Note, that the generation process for both networks only differs for the variable “Workinghours". For the train-sets we simulated 3500 datapoints, and for the validation sets and each of the 10 test sets we simulated 500 points.

$$\begin{aligned}
\text{Sex} &\sim \text{Bernoulli}(0.5) \\
\text{Redline}_{\text{Sex}=0} &\sim \mathcal{N}(170, 2) \\
\text{Redline}_{\text{Sex}=1} &\sim \mathcal{N}(175, 2) \\
\text{Skills} &\sim \mathcal{N}(5, 2) \\
\text{Workinghours} &\sim \mathcal{N}(5, 2) && \textit{non expl.disc.} \\
\text{Workinghours} &\sim 3 \times \text{Sex} + \mathcal{N}(5, 2) && \textit{expl.disc.} \\
\text{Wage} &\sim 3 \times \text{Skills} + 3 \times \text{Workinghours} + \mathcal{N}(5, 2) \\
\text{Wage Prob.} &= \frac{\text{Wage} - \min(\text{Wage})}{\max(\text{Wage}) - \min(\text{Wage})} \\
\text{Wage Labels} &= \text{Wage Prob.} > 0.5
\end{aligned}$$

make this simulation more realistic, only labels from women who neither have a high amount of working hours, nor a high level of skills is changed. To test the situation testing algorithms, we split each dataset into train- validation and 10 test sets. The validation set is used to choose the best value of k and t (and where applicable λ), while the test sets are used to evaluate the performance of the algorithms.

2.5.2 Experimental Setup

We tested the following algorithms to see which are best at predicting the non-biased class labels of the non-privileged instances (as given by the ground truth).

- **Baseline:** The discrimination scores of the baseline are based on a classifier trained on the privileged dataset. Assuming that no discrimination occurs in this part of the data, the trained classifier should predict “fair" labels on the non-privileged data as well. Thus the discrimination scores of a non-privileged instance with a negative decision label is taken as the classifier's predicted probability that their decision label should have been positive
- **Situation Testing k + k**: The situation testing algorithm as originally proposed by Luong, using one of the following distance measures:
  1. Luong - the distance function as given in section 2.2.1
  2. Zhang - the distance function as given in section 2.2.2
  3. Weighted Euclidean - the learned function as described in section 2.3

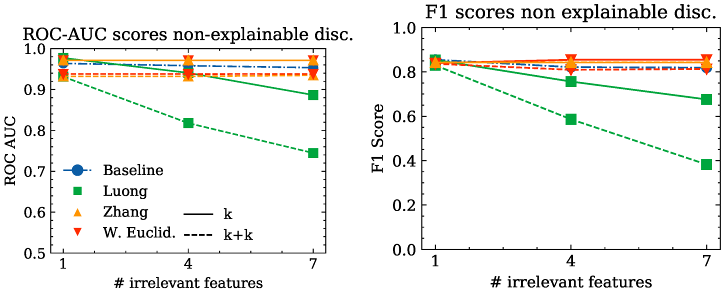
  Recall that with this algorithm, the discrimination scores of an instance are based on both its k unprivileged and k privileged neighbors

- **Situation Testing k**: The situation testing algorithm utilizing one of the three listed distance functions. This time discrimination scores are based on the positive decision rates among an instance's *k* privileged neighbors only.

As a performance measure, we use the AUC score of the ROC curve obtained on the discrimination scores. After turning the discrimination scores into discrimination labels (with the approach described in section 2.4.4), we also calculate the F1-score for each approach. All performance measures are calculated and then averaged over the 10 test sets.

2.5.3 Discrimination Detection without Explainable Discrimination

Looking at Figure 2.3a and 2.3b we see that in the non-explainable discrimination dataset with just one irrelevant attribute, utilizing the *k* approach results in similar performances of the situation testing algorithms with the different distance functions. All of them outperform the baseline, and interestingly all of them slightly outperform the $k + k$ approach utilizing the same distance functions.



(a) ROC AUC Score

(b) F1 Score

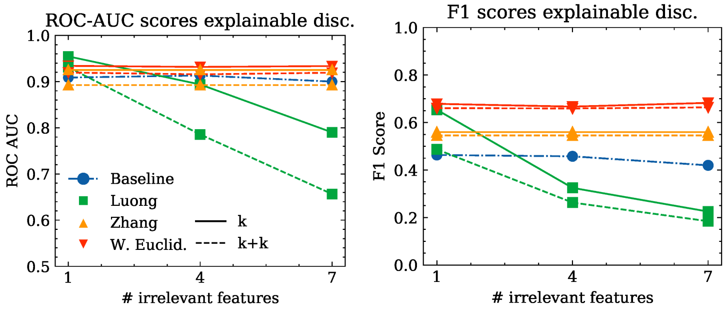
hyperparameters			
	k	λ	**t**
Baseline	-	-	0.5
Luong	10	-	0.25
Zhang	30	-	0.08
W. Euclid	10	0.07	0.25

Figure 2.3: Performance on non-explainable discrimination dataset. The situation testing algorithms using Luong's distance function, are the only ones dropping in performance when irrelevant attributes are added to the data.

This could be the case because the used number for *k* was not optimal for the $k + k$ approach. Additionally, the worse performance can be explained by the reasons listed in section 2.4.2: to know whether an individual of a non-privileged group was discriminated, we do not necessarily need the information about other members of the non-privileged group. When adding three or six additional irrelevant attributes to the dataset, we see how the situation testing algorithm utilizing Luong's distance function is the only one dropping in performance. This is the case because the distance function does not distinguish between the relevant and irrelevant attributes of the dataset. With one irrelevant attribute, this is not a problem, since it is still easy to find neighbors for an instance that are similar regarding all its features, even the ones that are not important for the decision problem. The more irrelevant attributes are added, the more "distraction" there is: with Luong's distance function we end up selecting neighbors for every instance that are similar on the irrelevant attributes, but not on those attributes that matter. The fact that the performance of the Weighted Euclidean's distance does not drop when more irrelevant attributes are added, shows that it has learned to disregard these attributes. Different than with Zhang's approach, this is accomplished without relying on the presence of a *CBN*.

2.5.4 Discrimination Detection in the Presence of Explainable Discrimination

Looking at Figures 2.4a and 2.4b we see that the results of the situation testing algorithms are a bit worse than in the previous dataset.



(a) ROC AUC Score

(b) F1 Score

hyperparameters			
	k	λ	t
Baseline	-	-	0.50
Luong	10	-	0.40
Zhang	10	-	0.01
W. Euclid	10	0.09	0.25

Figure 2.4: Performance on explainable discrimination dataset. The Weighted Euclidean distance outperforms the other distance functions. Again, the performance with Luong's distance function drops with more irrelevant attributes.

This is no surprise given that "Workinghours", an attribute relevant to the decision label, is distributed differently among men and women. If we try to assign a discrimination score to a woman with a low amount of workinghours it is hard to find similar male neighbours, which may lead to incorrect results. The performance of the baseline also is considerably worse than in the non-explainable discrimination dataset. To explain this, recall that its discrimination scores are based on a classifier's predicted class labels for women, after being trained on the data of men. During training, the classifier only saw instances with a high amount of workinghours, as men typically score high on this attribute. Thus it is logical that the classifier is not able to deal with the female dataset instances, which typically have a lower value for this feature. Due to the same reasons as described in section 2.5.3, we further see that for the situation testing algorithms the $k+k$ approach yields worse performances than the k approach, and that the performance of Luong's distance function drops when more irrelevant attributes are added to the data. Again, this effect does not occur when utilizing one of the more refined distance functions. This also shows that our distance function has correctly learned the difference between red-lining attributes and attributes like "Workinghours", which are correlated to the sensitive attribute but also in itself indicative for the decision label.

2.6 Qualitative Experiments on Real Data

To investigate the working of the new distance function on a more realistic dataset, we run some experiments on the "Adult" dataset, where for each individual is recorded whether their annual income is higher than $50k, along with information about their age, amount of working hours, education level, etc. As a sensitive attribute, we take the "Sex" of individuals, assuming that some women do not have a low income because of their specific characteristics but because of historical discrimination towards their sex. Since we do not know the true discrimination labels of the adult dataset, we will study the utility of the distance functions by exploring some case examples.

2.6.1 Case examples

We have visualized some properties of the 10 nearest unprivileged neighbours of two dataset instances, as selected by the distance functions. For each interval-scaled feature, we display the neighbors' mean value on this feature, while for ordinal-/and nominal scaled features we show the number of neighbors that have the same value as the instance in question on that feature (denoted by $|S|$).

Table 2.2: Some properties of the neighbors of two dataset instances, as selected by the distance functions. $|S|$ here denotes the number of neighbors that have the same value on the given feature as the instance in question.

instances			properties of neighbors as selected by distance functions		
	feature	**value**	**Luong**	**Zhang**	**W. Euclid.**
# 1	age	44	$\mu = 44.5$	$\mu = 45.9$	$\mu = 45.7$
	education level	Doctorate	$\|S\| = 6$	$\|S\| = 10$	$\|S\| = 10$
	hours per week	38	$\mu = 42.6$	$\mu = 55.9$	$\mu = 41.6$
	capital gain	0	$\mu = 0$	$\mu = 0$	$\mu = 0$
	capital loss	0	$\mu = 0$	$\mu = 0$	$\mu = 0$
	native country	US	$\|S\| = 10$	$\|S\| = 10$	$\|S\| = 10$
	marital status	Married	$\|S\| = 10$	$\|S\| = 10$	$\|S\| = 10$
	workclass	Governmental	$\|S\| = 10$	$\|S\| = 1$	$\|S\| = 4$
# 2	**feature**	**value**	**Luong**	**Zhang**	**W. Euclid.**
	age	42	$\mu = 41.5$	$\mu = 36.4$	$\mu = 35.6$
	education level	High School	$\|S\| = 10$	$\|S\| = 10$	$\|S\| = 10$
	hours per week	50	$\mu = 47.6$	$\mu = 40.2$	$\mu = 44.8$
	capital gain	5455	$\mu = 1956.4$	$\mu = 4202.3$	$\mu = 5092.1$
	capital loss	0	$\mu = 0$	$\mu = 0$	$\mu = 0$
	native country	US	$\|S\| = 10$	$\|S\| = 10$	$\|S\| = 9$
	marital status	Divorced	$\|S\| = 10$	$\|S\| = 3$	$\|S\| = 3$
	workclass	Private	$\|S\| = 10$	$\|S\| = 9$	$\|S\| = 9$

To understand these findings better we visualized the CBN learned for Zhang's distance function and the top 6 features ranked by their importance to the decision attribute, according to the Weighted Euclidean distance.

Coherent with what we see in Figure 2.5 and Table 2.6, Zhang's and the Weighted Euclidean distance function put a high emphasis on finding neighbours with a similar "capital gain" and "education level" as the instances in question. This comes at a cost on the similarity of features, like "marital status" that are less important for the decision attribute. Still, our learned distance function is better than Zhang's function at finding

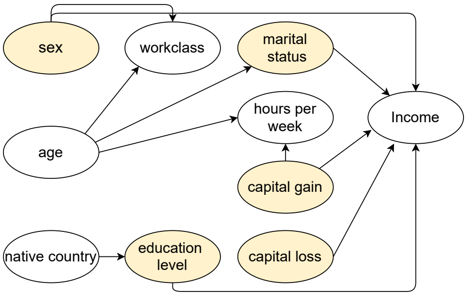


Figure 2.5: CBN for Zhang's distance function.

Rank	Feature Importance According to Weight. Euclid.
#1	education level
#2	capital gain
#3	capital loss
#4	age
#5	m. status = married
#6	hours per week

Figure 2.6: Ranked feature importances, according to learned Weighted Euclidean distance function

neighbours similar on attributes like "hours per week", which may not be most important for the decision task but also not trivial. This makes sense, given that Zhang's approach only defines distances based on the attributes directly affecting the decision attribute, and "hours per week" is not one of them. Looking at the neighbours selected with Luong's approach, we see that a lower similarity on specific features results in a higher similarity on others. Especially for instance #2 this might seem desirable, as here similarity is quite high for most features. However, seeing that the neighbors differ a lot on "capital gain", an attribute highly indicative for the decision label, this intuition is counteracted. Given that we are dealing with a complex dataset, where it is not always possible to find neighbors that are similar in all regards to some instance, it makes sense that a learned distance function priorities similarity on the most important features of the data. This especially holds, when recalling how the presence of irrelevant attributes worsened the performance of Luong's distance function in the simulated datasets.

2.7 Discussion and Conclusion

We have shown how a learned Weighted Euclidean distance function can be applied in the situation testing algorithm originally proposed by Luong [89], to find discrimination in data. The results on the simulated datasets show an advantage of utilizing the learned distance function over previously defined ones. This especially holds when a dataset contains multiple irrelevant attributes. With the experiments on a realistic dataset, the performance is less straightforward to assess, however, there are indications that our learned distance function also performs better here. Nevertheless, we here observed a downside of the situation testing methodology: the more features a dataset has, the more difficult it is to find neighbors for an instance that are similar on all relevant attributes. Consequentially, the discrimination scores based on these neighbors might not be very accurate. As seen in the experiments on the simulated data, the same problem occurs when a dataset contains explainable discrimination and the unprivileged and privileged group differ on an attribute relevant to the decision problem. Given these problems, we emphasize that the situation testing algorithm should merely be used as a tool to support discrimination detection. For instance, one could use the learned distance function to find the nearest neighbors of a potentially discriminated instance. Whether these neighbors are similar enough to base a discrimination judgement upon, should be decided by a human auditor.

In this study, we did not investigate yet how the situation algorithm with our proposed distance function, performs when it comes to discrimination prevention. The task of discrimination prevention is tackled by first using the algorithm to detect discrimination and then remove it, such that a classifier trained on this de-biased data does not learn to discriminate. How this approach affects the fairness and accuracy of a classifier, and how these performance measures compare to other fair learning algorithms, can be a direction for future research.

Chapter 3

Interpretable and Fair Selective Classification

In Chapters 1 and 2 of this thesis we addressed how to detect discrimination in ADM systems before they are deployed. While this is a crucial step towards making fairer ADM algorithms, it is also important to overlook their decision-making as they are being put to use. In this chapter, we will explore the potential of abstaining classifiers to achieve this goal. Traditionally, abstaining classifiers were designed to increase the accuracy of a decision process, by allowing them to refrain from making predictions in cases of uncertainty. Now we will see how this framework can be extended to reduce the discriminatory effect of a decision process, by letting classifiers abstain in cases of unfairness. Specifically, we introduce IFAC, an Interpretable and Fair Abstaining Classifier, that makes unfairness-based rejects based on inherently explainable fairness checks, namely rule-based approaches and the previously introduced method of situation testing. We show how by rejecting possibly unfair predictions, IFAC reduces demographic parities in errors and positive decision rates in the non-rejected data. Further, we illustrate with some examples of the explanations behind the rejections, how human auditors can be empowered to review rejected instances and make more well-informed decisions on them[1].

3.1 Introduction

The previous two chapters of this thesis have made a clear case for why detecting bias in an ADM system requires rigorous audits that go beyond single numerical measures of fairness. Efforts should be put into adapting to the specific context of a decision-making

[1] This chapter is based on: *Lenders, D., Pugnana, A., Pellungrini, R., Calders, T., Pedreschi, D. & Giannotti F. (In press). Interpretable and fair mechanisms for abstaining classifiers. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases 2024.* Daphne Lenders is the first author of this work, as she was responsible for conducting experiments and writing the majority of the work

system, as well as understanding where discrimination occurs on a more fine-grained individual level. It is clear that when it comes to the resolution of biases in an ADM system similar considerations need to be made: to ensure that a system does not discriminate against people throughout its deployment, we need a thorough understanding of which of its decisions are unfair and should be overridden. Reflecting this concern, there is a growing consensus that automated algorithms alone are not enough to successfully resolve bias, but instead should be actively overseen and adapted by human experts with sufficient knowledge about a domain and its historic biases. This call for human-in-the-loop approach is now even mandated by AI legislation, such as the EU AI Act [45].

A way to put humans in the loop during the deployment of a system is provided by the framework of selective classification. The original idea behind this framework is to build a classifier that abstains from making a prediction when it is not certain about it. Though this idea dates back to the 1970s [29], it has only barely been explored in the context of increasing the fairness of models, by abstaining from predictions that might be unfair. In this chapter, we propose a method to add this unfairness-based rejection mechanism to an abstaining classifier. Emphasizing the importance of measuring unfairness both through larger statistical patterns and local assessments, this mechanism is based on group and individual fairness metrics. Further, we ensure that all unfairness-based rejects are completely interpretable to a human auditor, such that the explanations behind the rejections can help them in reviewing the original predictions and override them where necessary. We name our methodology IFAC (Interpretable Fair Abstaining Classifier) and show how by making rejections both based on the uncertainty and unfairness of predictions, it increases accuracy and fairness over all non-rejected ones. By providing examples of the explanations behind its rejections, we highlight how human auditors could use them to make well-informed decisions about alternative (more fair) predictions.

3.2 Related Literature

Prediction with a Reject Option. The idea to allow a machine learning model to abstain in the prediction stage dates back to the 1970s, when it was introduced for classification tasks [29]. Two main frameworks allow one to learn abstaining models, i.e. ambiguity rejection and novelty rejection [60]. The former focuses on abstaining from instances where mistakes are more likely; the latter builds methods that abstain on instances that are largely dissimilar from the training data distribution [81, 102, 134]. Within ambiguity rejection, we can further distinguish between Learning to Reject (LtR) [29] and Selective Prediction (SP) [43]. The former (LtR) requires one to define a class-wise cost function that penalizes mispredictions and rejections [30, 33]. The latter (SP) requires instead one to either pre-define a target coverage c to achieve and minimize the risk *(bounded-abstention)* [57, 68, 105, 106], or fix a target risk e to guarantee and maximize the coverage *(bounded-improvement)* [55, 56].

Fairness and Reject Option. There are a few works that analyze the effects on fairness caused by a reject option. Jones et al. [71] show that even if abstaining can improve the overall accuracy, some demographic groups can be negatively impacted by the reject option. Lee et al. [85] propose a surrogate loss for the classification task considering performance on different subgroups of instances. The proposed loss allows enforcing a sufficiency condition to avoid unfair results. A similar approach for the regression task is proposed by Shah et al. [119]. Schreuder & Chzhen [117] provide a theoretical analysis of the selective classification framework when introducing a fairness constraint in the bounded-abstention problem.

Explainability and Reject Option. The study of explainable AI (XAI) methods in the context of abstaining classifiers is limited. Fischer et al. [50] propose a reject option for natively interpretable models such as prototype-based ones. Artelt et al. [4] consider counterfactual techniques to explain reject options of learning vector quantization classifiers. Artelt & Hammer [5] introduce semi-factual explanations for the reject option, yielding a model-agnostic approach at the expense of potentially high complexity. Finally, Artelt et al. [6] propose a model-agnostic framework to explain the abstention mechanism, including counterfactual, semi-factual, and factual approaches.

3.3 Background

3.3.1 Selective Classification

Consider the triplet $(\mathbf{A}, \mathbf{G}, Y)$: $\mathbf{G}$ represents the legally-grounded features and takes values in $\mathcal{G} \subseteq \mathbb{R}^{d_g}$; $\mathbf{A}$ refers to the sensitive attributes and takes values in $\mathcal{A} \subseteq \mathbb{R}^{d_a}$; Y is the (binary) target variable, whose domain is $\mathcal{Y} = \{0, 1\}$. For example, if Y encodes being rich and our goal is to predict Y given some set of features, $\mathbf{G}$ could include educational level and employment status, while $\mathbf{A}$ could refer to gender or race. We denote with $\mathcal{X} = \mathcal{G} \times \mathcal{A}$ the whole feature space and with $\mathbf{X} = (\mathbf{G}, \mathbf{A})$ the pair of both legally grounded and sensitive features.

Given the hypothesis space $\mathcal{H}$ of functions (classification models) mapping $\mathcal{X}$ to $\mathcal{Y}$, a learning algorithm aims to find a hypothesis $h \in \mathcal{H}$ such that it minimizes some risk measure $R(h) = \mathbb{E}[l(h(\mathbf{X}), Y)]$, where $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a *loss function* and $\mathbb{E}$ is computed over the joint probability distribution $P(\mathbf{X}, Y)$.

To reduce the classifier's error rates, one can add a selection mechanism that allows the model to abstain from predicting over more difficult-to-classify instances. More formally, we can define a selective classifier[2] as:

$$(h, g)(\mathbf{x}) = \begin{cases} h(\mathbf{x}) & \text{if} \quad g(\mathbf{x}) = 1 \\ \text{abstain} & \text{otherwise,} \end{cases} \tag{3.1}$$

where $g : \mathcal{X} \to \{0, 1\}$ is the so-called *selection function* or *rejector*[3].

In practice, the selection function is often obtained by setting a threshold τ on a confidence function $v : \mathcal{X} \to \mathbb{R}$, which determines the portion of the data on which the classifier is more likely to misclassify. In such a case, the selection function can be defined as $g(\mathbf{x}) = \mathbb{1}\{v(\mathbf{x}) \geq \tau\}$.

To avoid rejecting too many instances, the selective classification framework introduces *the coverage*, i.e. the percentage of instances for which the selective classifier must provide

[2] In this work, we use the terms *abstaining* and *selective* interchangeably.

[3] We use the term abstain and reject when $g(\mathbf{x}) = 0$ and accept or selects when $g(\mathbf{x}) = 1$.

a prediction. Coverage is denoted as $\phi(g) = \mathbb{E}[g(\mathbf{X})]$ and can be traded off for performance improvements. In this case, performance is measured through the risk over the accepted region, commonly called the *selective risk* and defined as $R(h, g) = \frac{\mathbb{E}[l(h(\mathbf{X}),Y)g(\mathbf{X})]}{\phi(g)}$.

To find a selective classifier that minimizes selective risk, it is necessary to select a lower bound c as a *target coverage* [57]. Given a target coverage c, an optimal selective predictor (h, g) (parameterized by θ^*, ψ^*) is defined as:

$$\underset{\theta\in\Theta,\psi\in\Psi}{\arg\min}\, R(h_\theta, g_\psi) \quad \text{s.t.} \quad \phi(g_\psi) \geq c \tag{3.2}$$

We learn the optimal parameters using an empirical counterpart of selective risk and coverage, using an i.i.d. dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ drawn from P.

Finally, we call *coverage-calibration* the post-training procedure of estimating the threshold τ for the target coverage c specified in Eq. 3.2. This is generally done by estimating the $(1-c)\cdot 100$-th percentile of the confidence function over a held-out calibration dataset.

3.3.2 Measuring Fairness With Association Rules & Situation Testing

Association Rules: In our methodology, we make use of association rules to identify discriminatory behaviour of a base classifier h, upon which g can decide to reject its predictions. Let us assume we have access to a dataset of realizations $\mathcal{D}$. We recall $\mathbf{x}_i = (\mathbf{g}_i, \mathbf{a}_i) = (g_i^1, \cdots, g_i^{d_g}, a_i^1, \cdots, a_i^{d_a})$, where g_i^j refers to the value taken by the j^{th} legally grounded feature of instance i and a_i^j to the j^{th} sensitive feature of instance i.

We call a specific realization of a single variable within $\mathbf{x}_i$ an *item*, e.g. if we consider the variable `race`, `race=White` is an item. Let $\mathcal{I}$ be the set of all possible items. A subset I of $\mathcal{I}$ is called an *itemset*.

We can decompose I into its legally grounded and sensitive parts, $I = (I_G, I_A)$, where I_G is an itemset containing only legally grounded features and I_A is an itemset that contains only sensitive ones. A transaction T is a subset of I with exactly one item for every feature in $\mathbf{x}$. In other words, a sampled instance's features $\mathbf{x}_i$ can be seen as a transaction T. For a transaction T, we say T *verifies* itemset (I_G, I_A) if $(I_G, I_A) \subseteq T$. The support of itemset (I_G, I_A) with respect to the dataset $\mathcal{D}$ is denoted as $supp_{\mathcal{D}}((I_G, I_A)) = \frac{|\{T\in\mathcal{D}:(I_G,I_A)\subseteq T\}|}{|\mathcal{D}|}$.

A decision rule is an expression $(I_G, I_A) \longrightarrow Y$. The support of a decision rule is $supp_{\mathcal{D}}((I_G, I_A) \rightarrow Y) = supp_{\mathcal{D}}((I_G, I_A), Y)$. The confidence of the rule is then defined as $conf_{\mathcal{D}}((I_G, I_A) \rightarrow Y) = \frac{supp_{\mathcal{D}}((I_G,I_A),Y)}{supp_{\mathcal{D}}((I_G,I_A))}$.

To measure the impact of the sensitive features of a decision rule, the Selective Lift (*slift*) measure introduced by Pedreschi et al. [101] can be used. In this chapter we use the definition *by difference* of *slift*, which is detailed as follows:

$$slift_{\mathcal{D}}((I_G, I_A) \rightarrow Y) = conf_{\mathcal{D}}((I_G, I_A) \rightarrow Y) - conf_{\mathcal{D}}((I_G, \neg I_A) \rightarrow Y) \tag{3.3}$$

Computing $conf_{\mathcal{D}}(I_G, \neg I_A) \longrightarrow Y$ requires one to take the confidence of all the transactions that verify I_G but do not verify I_A.

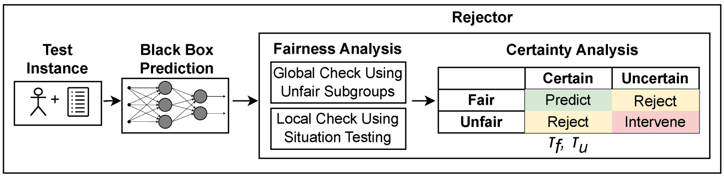
Example. Consider an association rule `race = Black, education = Masters` $\longrightarrow$ `income = low`, with `race` $\subseteq$ **A** and `education` $\subseteq$ **G** and `income` $= Y$. Imagine the confidence of this rule is 0.90 and its slift is 0.50. This means that the confidence of `race` $\neq$ `Black, education = Masters` $\longrightarrow$ `income = low` is 0.90-0.50 = 0.40. Because of this high difference `race = Black, education = Masters` could be seen as a subgroup at risk of discrimination.

As indicated by Pedreschi et al. [100], decision rules can be learned on the original data using algorithms like Apriori [2] and then filtered according to fairness-based policies.

Situation Testing: Since association rules only detect global discrimination patterns, one can use the Situation Testing algorithm to further analyse fairness on a local level [89]: To check whether instance $\mathbf{x}_i$ receives a fair outcome Y, we use a distance function to search $\mathcal{D}$ for $\mathbf{x}_i$'s k-nearest neighbors from a reference group and a non-reference group, meaning we obtain two sets of instances $\mathcal{K}^r_{tr}$ and $\mathcal{K}^{nr}_{tr}$. A reference group is defined by sensitive feature values of those instances from the data we assume to be treated favorably, for instance (`race = White, sex = Male`). All instances not belonging to this group are seen as the non-reference group. To define instance $\mathbf{x}_i$'s individual discrimination score we calculate the ratio of positive decision ratio for $\mathcal{K}^r_{tr}$ and $\mathcal{K}^{nr}_{tr}$: $dec_r = \frac{|\{j \in \mathcal{K}^r_{tr} : y_j = 1\}|}{k}$, $dec_{nr} = \frac{|\{j \in \mathcal{K}^{nr}_{tr} : y_j = 1\}|}{k}$ and take the difference between both $(dec_r - dec_{nr})$. If this score exceeds some individual discrimination threshold t, it indicates that the treatment reserved to instance i depends on its sensitive characteristics.

3.4 Methodology

We propose to learn a selective classifier that does not only reject instances based on the uncertainty of their predictions but also their unfairness. In doing so we can decrease unfairness over all non-rejected instances. Further, by providing explanations for why some predictions are marked as unfair, we aid human reviewers in understanding whether the fairness concerns are indeed justified and enable a more informed decision process over them. We call our approach IFAC (Interpretable and Fair Abstaining Classifier). The intuition behind IFAC is visualized in Figure 3.1: on top of the base classifier h we have our rejector g, which takes an instance's features $\mathbf{x}_i$ and the classifier h's prediction as its input. The rejector first executes a global fairness analysis on this instance, checking if it falls under any subgroups at risk of discrimination, as identified by discriminatory association rules (section 3.3.2). If it does, it performs a local fairness check using Situation Testing [89], evaluating how the prediction for $h(\mathbf{x}_i)$ compares to the labels of similar instances in the data. After this, a *certainty assessment* is performed. Depending on the outcome of the assessment and the former fairness analysis there are four possibilities for our rejector: in case the prediction is deemed as fair and it exceeds a dedicated confidence threshold, the prediction is kept. Contrary, fair predictions that fall below this threshold are rejected. If we are dealing with an unfair prediction exceeding a separate confidence threshold for unfair data, it also gets rejected: though the prediction is certain, we have reasons to doubt it, because it is unfair. Finally, on predictions that are both unfair and



	Certain	Uncertain
Fair	Predict	Reject
Unfair	Reject	Intervene

Figure 3.1: Intuition behind IFAC

uncertain, IFAC flips the original classifier $h(\mathbf{x}_i)$ prediction. The reasoning behind these interventions is that predictions that are neither fair nor certain are probably inaccurate, to begin with, and it is safe to alter them. A complete walk-through example of how IFAC makes rejections is provided in Appendix A.1.

Now that we have described the basic intuition behind how IFAC is applied, we outline how it is learned. Given some data $\mathcal{D}$, we split it into a training set $\mathcal{D}_{tr}$ and two validation sets $\mathcal{D}_{val_1}$, $\mathcal{D}_{val_2}$. Then, given the target coverage c and the unfair reject weight w_u[4], IFAC is devised as follows:

1. **Learn a classifier:** we train classifier h from $\mathcal{D}_{tr}$. We highlight that any off-the-shelf probabilistic classifier can be considered, making our approach model-agnostic;
2. **Learn at-risk subgroups:** we extract association rules from validation set $\mathcal{D}_{val_1}$. The rules allow us to understand if there are correlations between sensitive features $\mathbf{S}$ and predictions of h, and, consequently, identify at-risk subgroups [100];
3. **Situation Testing:** we prepare the hyperparameters and distance function to run Situation Testing.
4. **Calibration:** we use the second validation set $\mathcal{D}_{val_2}$ to calibrate the rejection strategy, considering both *unfairness* and *uncertainty*:
   (i) the learned association rules are applied on $\mathcal{D}_{val_2}$;
   (ii) situation testing is performed for those instances falling under discriminatory patterns. This allows one to split the sample into a *fair* part $\mathcal{D}_{val_2^f}$ and an *unfair* one $\mathcal{D}_{val_2^u}$;
   (iii) depending on c and w_u, we estimate two different rejection thresholds, i.e. τ_f and τ_u. These thresholds are computed following the *coverage-calibration* procedure described in section 3.3, ranking instances w.r.t. the confidence function over samples $\mathcal{D}_{val_2^f}$ and $\mathcal{D}_{val_2^u}$ respectively.

Figure 3.2 summarizes the steps needed to learn IFAC. In the rest of this section, we further detail steps 2, 3, and 4.

3.4.1 Step 2: Learn At-Risk Subgroups

To learn global patterns of unfairness, we use discriminatory association rules, as described in section 3.3.2. To do so we apply h on the first validation set $\mathcal{D}_{val_1}$ and

[4] The unfair reject weight w_u determines how many rejections can be made based on unfairness concerns.

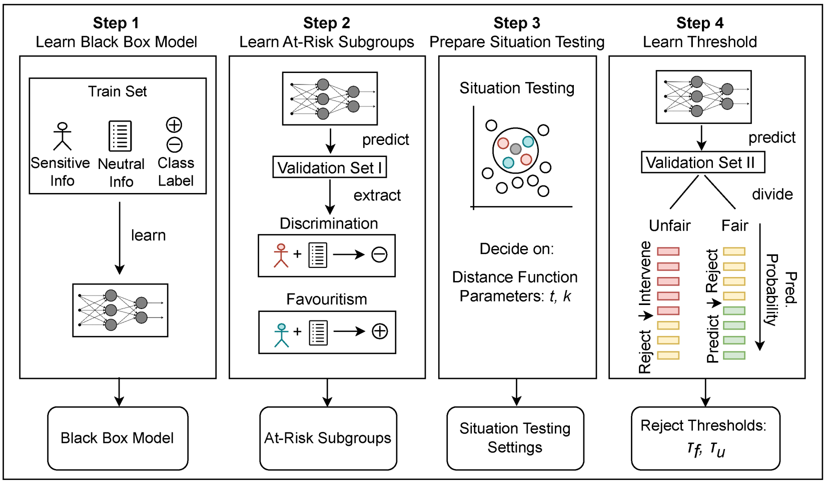


Figure 3.2: The four steps for learning IFAC.

extract the association rules for the data and hs predictions with the apriori algorithm. We do so separately for each sensitive feature value and their combination. For example, let us have two sensitive attributes `sex` and `race` with two possible values, `F,M` and `W,B` respectively. We apply apriori and extract rules for each of the itemsets: `{sex=M}`, `{sex=F}`, `{race=W}`, `{race=B}`, `{sex=M ∧ race=B}`, `{sex=M ∧ race=W}`, `{sex=F ∧ race=W}`, `{sex=F ∧ race=B}`. Thus, the number of rules found meeting minimum support is not biased towards the largest demographic groups in the data.

As per our previous notation, we extract rules in the form of $(I_G, I_A) \longrightarrow Y$, for some prediction outcome $h(\mathbf{x})$ in a binary classification setting $Y \in \mathcal{Y} = \{0, 1\}$. We say that rules with $Y = 0$ describe potentially discriminated subgroups, while rules with $Y = 1$ describe potentially favored ones. We extract favoring associations only for fixed reference groups defined for our data, e.g. white men (as described in section 3.3.2). After extracting both favoring and discriminatory associations, we filter out statistically significant rules meeting an *slift* threshold. We calculate statistical significance using Z-test, testing if the proportion of some decision outcome Y is significantly different for the groups (I_G, I_A) and $(I_G, \neg I_A)$ [24]. We only select rules with $p < 0.01$. Further, we filter out *high-slift* rules by checking for which ones the following holds:

$$conf_{\mathcal{D}_{val_1}}((I_G, I_A) \longrightarrow Y_v) - slift_{\mathcal{D}_{val_1}}((I_G, I_A) \longrightarrow Y_v) < 0.5 \tag{3.4}$$

Which in the context of binary classification is true *iff*:

$$conf_{\mathcal{D}_{val_1}}((I_L, \neg I_S) \longrightarrow Y_v) < conf_{\mathcal{D}_{val_1}}((I_L, \neg I_S) \longrightarrow \neg Y_v) \tag{3.5}$$

Intuitively, this means that we only select the groups $\{I_G, I_A\}$ for which negating the sensitive part of the group $(\{I_G, \neg I_A\})$ yields higher confidence for value Y_v w.r.t. the opposite value $\neg Y_v$ (brief proof in Appendix B.2).

3.4.2 Step 3: Situation Testing

Part of the abstention mechanism of IFAC is based on a local fairness check for instances that are covered by global discrimination patterns. The aim is to use the global check to identify larger subgroups at risk of unfair treatment, while the local check allows us to execute a more fine-grained analysis taking all of an instance's characteristics into account. Our local fairness check is performed via Situation Testing, comparing a prediction $h(\mathbf{x}_i)$ for instance i with the decision labels of similar instances from $\mathcal{D}_{tr}$ (see section 3.3.2). For the algorithm, a suitable distance function must be chosen e.g. by automatically learning one from the data [87] (see Chapter 2 of this thesis).

3.4.3 Step 4: Calibrate Rejection Strategy

Whether the rejector keeps, rejects, or intervenes on the original prediction for $\mathbf{x}$, depends on the (un)certainty of the base classifier. To evaluate the confidence of the classifier, we resort to the softmax response $v(\mathbf{x}) = \max_{y \in \mathcal{Y}} s_y$ [53,56], where $s_y(\mathbf{x}) \approx P(Y = y|\mathbf{X} = \mathbf{x})$ is an estimate of the conditional probability. We then estimate two thresholds τ_f and τ_u to choose between prediction, intervention, and abstention. The final selective classifier is in the form:

$$(h, g)(\mathbf{x}) = \begin{cases} h(\mathbf{x}) & \text{if } Fair(\mathbf{x}) \text{ and } v(\mathbf{x}) => \tau_f \\ \text{abstain} & \text{if } Fair(\mathbf{x}) \text{ and } v(\mathbf{x}) < \tau_f \\ 1 - h(\mathbf{x}) & \text{if } \neg Fair(\mathbf{x}) \text{ and } v(\mathbf{x}) < \tau_u \\ \text{abstain} & \text{if } \neg Fair(\mathbf{x}) \text{ and } v(\mathbf{x}) >= \tau_u \end{cases}$$

To learn τ_f and τ_u, h is applied on our second validation dataset $\mathcal{D}_{val_2}$ and its predictions are extracted. We then first extract those predictions that fall under discriminatory associations as learned in Step 2. After, we apply the Situation Testing algorithm as set up in Step 3 on those instances, and extract all that fail this individual fairness test. We consider those as the unfair fraction of the validation data ($\mathcal{D}_{val_2^u}$) and the remaining ones as the fair fraction $\mathcal{D}_{val_2^f}$. The number of rejections that can be made for both groups is determined by two parameters given by the user, namely the target coverage c and the unfair reject weight w_u. Given that the $\mathcal{D}_{val_2}$ consists of N instances of which N_u belong to $\mathcal{D}_{val_2^u}$ and N_f belong to $\mathcal{D}_{val_2^f}$, we calculate the number of total rejections ($\mathrm{N_{rej}}$), the number of unfairness-based rejections ($\mathrm{N_{ufr}}$) and the number of uncertainty-based rejections ($\mathrm{N_{ucr}}$) as follows:

$$\mathrm{N_{rej}} = \lceil (1 - c) \cdot \mathrm{N} \rceil; \quad \mathrm{N_{ufr}} = min(\lceil \mathrm{N_{rej}} \cdot w_u \rceil, \mathrm{N_u}); \quad \mathrm{N_{ucr}} = \mathrm{N_{rej}} - \mathrm{N_{ufr}} \tag{3.6}$$

We then proceed by separately ordering the fair and unfair instances of the validation data according to the confidence function $v(\mathbf{x})$. On the fair instances, we determine the threshold τ_f such that $\mathrm{N_{ucr}}$ instances fall below this threshold, and on the unfair sample such that $\mathrm{N_{ufr}}$ instances exceed τ_u.

3.5 Experimental Evaluation

The goal of our experimental section aims to address the following questions:

Q1: Does IFAC achieve comparable results to state-of-the-art selective classifiers in terms of predictive performance and fairness?

Q2: How does IFAC explain the drivers behind unfairness-based rejections, and how could these explanations be utilized?

Q3: How do *coverage* c and the *unfair-reject weight* u_w affect our results?

3.5.1 Experimental Settings

Data and Baselines. We run experiments considering two real datasets, namely ACSIncome [40] and WisconsinRecidivism [7]. The former is about predicting high or low income based on instances' education, occupation etc. We define `sex` (male vs. female) and `race` (white vs. black vs. other) as sensitive attributes and take the group of white men as our reference group. We compare their treatment to each intersectional group based on race and sex.

WisconsinRecidivism contains information about criminal defendants, like their type of offense, number of prior offenses, etc. The task is to predict if they will not recidivate. We take `race` as the sensitive attribute (white vs. black vs. other). Because of a base classifiers' lower False Negative and higher False Positive rates on white people, we define this as the reference group [5].

We use different classification algorithms, namely a Random Forest, a Neural Network, and an XGBoost Classifier. We fitted all models with the default parameters of the corresponding `Python` libraries. Starting from these base classifiers, we compare IFAC with the following model-agnostic methods:

- *Full Coverage* (FC): the classifier itself when predicting on all the instances ($c = 1.00$)
- *Uncertainty Based Abstaining Classifier* (UBAC): The plug-in algorithm by Herbei & Wegkamp [62]. This is the most well-known model-agnostic method and achieves state-of-the-art performance [104]. As for IFAC, we consider $v(\mathbf{x}) = \max_{y \in \mathcal{Y}} s_y(\mathbf{x})$ as the confidence function. The rejection threshold is computed according to the *coverage-calibration* procedure.

Because we consider discrimination based on non-binary sensitive attributes (and in the case of ACSIncome even intersectional discrimination), we do not compare with the fair abstention mechanism of Schreuder et al. [117] as a baseline, which only works on a single binary sensitive feature.

Hyperparameters. For **Q1** and **Q2**, we set $c = .80$ for the abstaining classifiers. Further, for IFAC we set the *unfair reject weight* (w_u) equal to 1.0. The intuition behind this is that

[5] For full details on the preprocessing steps executed on both datasets we refer to our github repository

Table 3.1: Performance Results ACSIncome and WisconsinRecidivism

		ACSIncome			WisconsinRecidivism		
		Acc.	**Rec.**	**Prec.**	**Acc.**	**Rec.**	**Prec.**
RF	FC	.78 ± .01	.57 ± .02	.65 ± .03	.62±.01	.77±.01	.65±.01
	UBAC	**.83** ± .01	**.62** ± .02	**.69** ± .03	**.65**±.01	**.83**±.01	**.66**±.01
	IFAC	.80 ± .01	.59 ± .04	.64 ± .03	**.65**±.01	**.83**±.01	**.66**±.01
NN	FC	.80 ± .01	.58 ± .03	.71 ± .03	.63±.01	0.74±.01	.65±.01
	UBAC	**.86** ± .01	**.62** ± .03	**.77** ± .03	**.66**±.02	**.77**±.01	**.68**±.02
	IFAC	.83 ± .01	.58 ± .03	.73 ± .02	**.66**±.02	.76±.01	**.68**±.02
XGB	FC	.81 ± .01	.60 ± .03	.73 ± .03	.63±.01	.77±.01	.65±.01
	UBAC	**.87** ± .01	**.64** ± .03	**.78** ± .03	**.66**±.01	**.83**±.01	**.68**±.01
	IFAC	.84 ± .01	.59 ± .03	.75 ± .03	**.66**±.01	.82±.01	**.68**±.01

if the coverage is large enough, IFAC should abstain from predicting any unfair instance, and only if not, fairness interventions should be performed. For the Situation Testing algorithm used by IFAC we set k, i.e. the number of neighbors used for the fairness comparisons to 10, and t to 0.3. For extracting discriminatory association rules we use the apriori algorithm of `apyori` with min. support of 0.01 and min. confidence of 0.85.

Metrics. For **Q1**, we evaluate predictive performance in terms of accuracy, precision, and recall on all non-rejected instances. Concerning fairness measures, we report the False Negative, False Positive, and Positive Decision Rates for the different demographic groups of each dataset. Further, we report the range and the standard deviation across demographic groups over these measures. Note, that we define these measures regarding the desirable label of each dataset. Hence, the positive decision ratio for ACSIncome is the ratio of *high* income prediction, and for WisconsinRecidivism it is the ratio of *non-recidivism* predictions.

Experimental Setup. We split each dataset into training, two validation, and a test part (40% for train, 15% for each validation, and 30% for test) and train the classifiers on the former. For IFAC we learn the discriminatory associations on the first validation set. The reject thresholds for both IFAC and UBAC are calibrated based on the second. Finally, we randomly split the test set into 10 samples [87] and compute the final metrics on each of these samples. We provide results as averages and standard errors over these 10 test set samples.

3.5.2 Results

3.5.2.1 Q1: Performance & Fairness

We describe the predictive performance on each dataset and each classifier-methodology combination in Table 3.1. As can be seen, both selective classification methods improve upon the performance of FC, however, for UBAC this improvement is slightly larger, especially for the income prediction task.

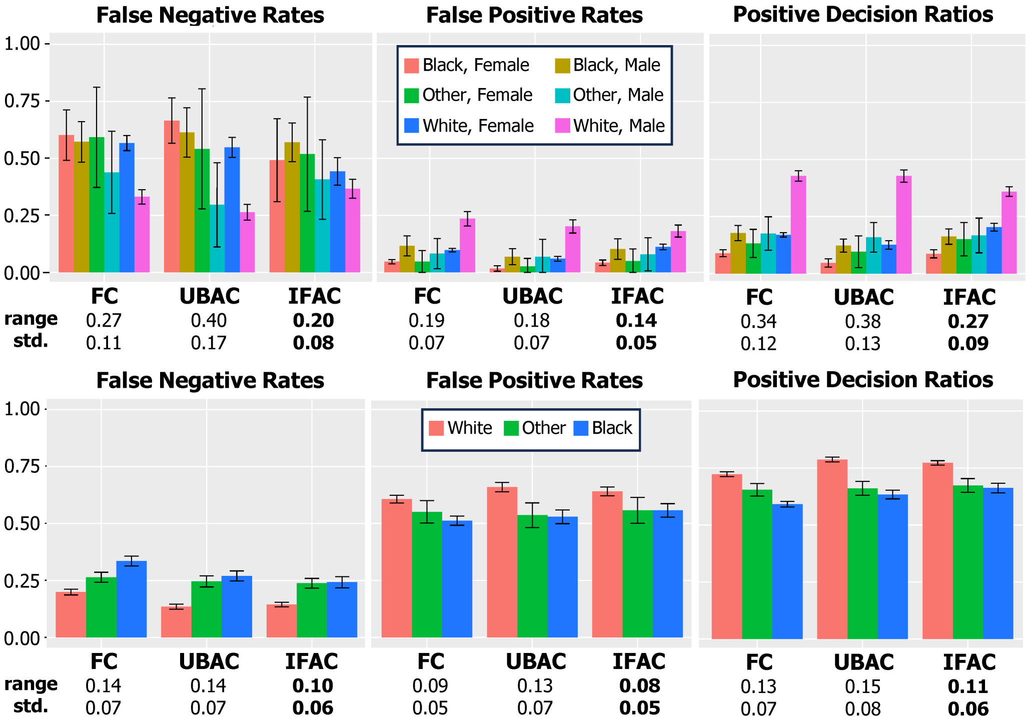
In Figure 3.3 we can see how the increased performance of UBAC comes at the cost of its fairness. In this Figure, we highlight the results of a Random Forest classifier combined with different selective classification methods, showing the average False Negative -, False Positive, and Positive Decision Rates (FNR, FPR, and PDR) over demographic groups (the results for Neural Networks and XGBoost follow the same patterns and are included in the Appendix A.3). We also highlight the range of these metrics across demographics (i.e. the performance difference between the highest- and lowest performing group) and the standard deviation. Fairer classifiers should score lower on both metrics, to ensure that there are no big performance differences across groups.

Starting with ACSIncome, we see that for UBAC this is not the case: we observe an especially unequal distribution of FNR across demographic groups, with the highest difference being 0.4 (between white men and black women). This difference is even higher than for the FC classifier, as the UBAC selection mechanism only decreases the FNR for white men while increasing it for others. With using IFAC this effect does not occur: through rejecting predictions that are at high risk of unfairness, FNRs decrease for minority groups like women or black people, and overall the rates become more equal across demographics, bringing the range down to 0.2 and the std. to 0.08. The patterns are slightly less strong when considering the FPR and PDR across demographics, but still hold. Similar patterns occur for WisconsinRecidivism: the range and standard deviation for FNR, FPR, and PDR across demographics decrease when using IFAC, while they increase with UBAC. We acknowledge that the effect is less strong here, but attribute this to IFACs selection criteria for unfair instances being too strict. In Appendix A.4 we show results with a lower threshold t for situation testing (meaning that more instances can get rejected out of unfairness concern), where IFAC makes FNR, FPR, and PDR nearly equal across groups. Further, we highlight how equalizing error rates across demographics is only the first step towards improving the fairness of the decision task. As we illustrate in the next section, enabling humans to review rejected instances and the explanation behind them, is the most crucial contribution of our method.

3.5.2.2 Q2: Explaining Unfair Rejections.

One of the main advantages of IFAC is that it can explain why rejected predictions are seen as unfair. In Figure 3.4 we show some explanations behind rejected instances for both of our datasets, and we use the ACSIncome case to highlight how a human expert can utilize them. We see two instances that were both rejected based on the same global pattern of unfairness: the classifier predicting "low income" ratios for black women, aged between 30 and 39 working in management, than for people with the same age and occupation, but different demographics. While an algorithm only analyses such patterns statistically, human experts can examine them with sensitivity surrounding their historical context. For instance, it is well known that racism and sexism contribute to hostile work environments for black women. Hence, a human expert can reason how these dynamics may hinder fair compensation in roles like management, that are normally associated with high salaries.

The results of situation testing provide further insight into the unfairness of the classifier: For both instances, a high ratio of the 10 most similar white men have a high income; explaining why their own low income predictions are marked as unfair. However, for the first instance, many of the white men considered for the comparison have a higher education level and amount of working hours than her. Since it makes sense, that people working part-time do not get the same compensation as people working full-time, the



	False Negative Rates			False Positive Rates			Positive Decision Ratios		
	FC	UBAC	IFAC	FC	UBAC	IFAC	FC	UBAC	IFAC
range	0.27	0.40	**0.20**	0.19	0.18	**0.14**	0.34	0.38	**0.27**
std.	0.11	0.17	**0.08**	0.07	0.07	**0.05**	0.12	0.13	**0.09**

	False Negative Rates			False Positive Rates			Positive Decision Ratios		
	FC	UBAC	IFAC	FC	UBAC	IFAC	FC	UBAC	IFAC
range	0.14	0.14	**0.10**	0.09	0.13	**0.08**	0.13	0.15	**0.11**
std.	0.07	0.07	**0.06**	0.05	0.07	**0.05**	0.07	0.08	**0.06**

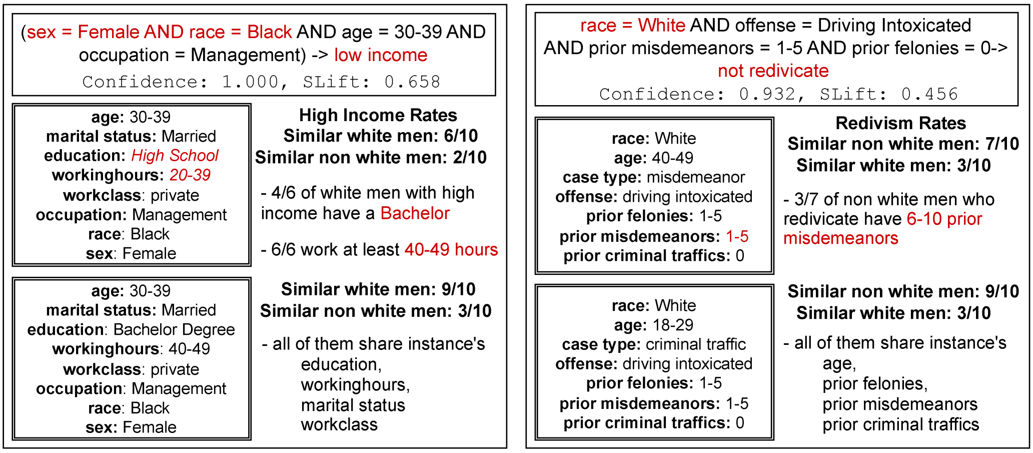
Figure 3.3: Performance measures over demographic groups when applying a Random Forest in combination with various selective classifiers on ACSIncome (above) and WisconsinRecidivism (below). A regular UBAC increases differences in error- as well as positive decision rates among groups. Using IFAC, and rejecting instances based on unfairness, diminishes these differences.

low income prediction could be seen as justified and a human reviewer could decide to keep it. For the second case, all similar white men do share the instances' education level, working hours, etc. Hence, there is no justification for why she would be the only one receiving a low income prediction, and a human expert could decide to override this decision.

To conclude, these examples show how IFAC's interpretable-by-design rejector can have a large impact in increasing the fairness of a decision process. In particular, our approach goes beyond a rough statistical analysis of discriminatory patterns and allow for the integration of human domain knowledge to achieve a much deeper fairness assessment.

3.5.2.3 Q3: Effects of c and w_u.

In this section, we explore the effect of parameters c and w_u on IFAC's performance. Out of space constraints, we only report the results with a Random Forest as a base-classifier on ACSIncome. The results for the other classifiers and the other dataset follow the same pattern and are included in Appendix A.5. In Figure 3.5 we visualize how the accuracy, the range in positive decision ratio across demographics, and the standard deviation



(sex = Female AND race = Black AND age = 30-39 AND occupation = Management) -> low income
Confidence: 1.000, SLift: 0.658

age: 30-39
marital status: Married
education: *High School*
workinghours: *20-39*
workclass: private
occupation: Management
race: Black
sex: Female

High Income Rates
Similar white men: 6/10
Similar non white men: 2/10

- 4/6 of white men with high income have a Bachelor
- 6/6 work at least 40-49 hours

age: 30-39
marital status: Married
education: Bachelor Degree
workinghours: 40-49
workclass: private
occupation: Management
race: Black
sex: Female

Similar white men: 9/10
Similar non white men: 3/10

- all of them share instance's education, workinghours, marital status workclass

race = White AND offense = Driving Intoxicated AND prior misdemeanors = 1-5 AND prior felonies = 0-> not redivicate
Confidence: 0.932, SLift: 0.456

race: White
age: 40-49
case type: misdemeanor
offense: driving intoxicated
prior felonies: 1-5
prior misdemeanors: 1-5
prior criminal traffics: 0

Redivism Rates
Similar non white men: 7/10
Similar white men: 3/10

- 3/7 of non white men who redivicate have 6-10 prior misdemeanors

race: White
age: 18-29
case type: criminal traffic
offense: driving intoxicated
prior felonies: 1-5
prior misdemeanors: 1-5
prior criminal traffics: 0

Similar non white men: 9/10
Similar white men: 3/10

- all of them share instance's age, prior felonies, prior misdemeanors prior criminal traffics

Figure 3.4: Examples for ACSIncome (left) and WisconsinRecidivism (right) of two rejected instances, and the explanation behind their rejections.

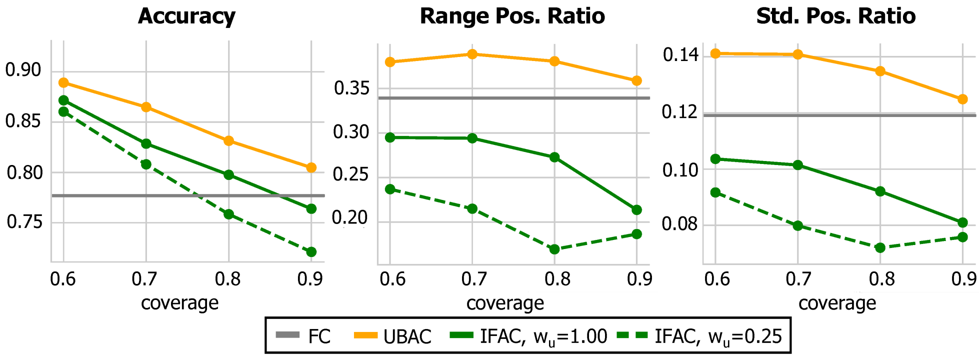


Figure 3.5: Effects of c and w_u parameters in our selective classification settings.

change as a function of the coverage and the w_u. Unsurprisingly, for both UBAC and IFAC the accuracy drops as the coverage increases. Regardless of the coverage and the w_u UBAC outperforms IFAC. Further, we see that a lower w_u comes at the cost of accuracy, especially when the coverage is high. Intuitively this makes sense: w_u determines how many of the unfair predictions are rejected, and for how many an intervention is performed. With the low weight of 0.25, the majority of unfair prediction labels are simply flipped, and only the ones with very high prediction probability are abstained from. With an increase in coverage, this pattern is more extreme, as the general number of instances that can be abstained from is lower. When observing the effect of differing coverages and w_u on the fairness of the predictions, we observe that performing more interventions (as a result of a lower w_u) has a desirable effect: both the range and standard deviation of positive decision ratios decreases across demographics. The effect is again larger for higher coverages because fewer allowed rejections mean more interventions, which bring the positive decision ratios across demographic groups closer together.

3.6 Discussion & Conclusion

In this chapter, we have introduced IFAC, an Interpretable and Fair Abstaining Classifier. This classifier rejects predictions from a base classifier, both in cases of uncertainty and unfairness. Unfairness rejections are based on the interpretable-by-design methods of unfair association patterns and situation testing. Through our experiments, we have shown how using our abstention mechanism yields satisfying overall performance, while improving fairness across demographic groups over all non-rejection instances. This stands in contrast to a regular uncertainty-based abstaining classifier, that does not take the fairness of predictions into account. We have also shown how the explanations behind our abstention mechanism, can empower human decision-makers to review the rejected instances and make fairer decisions for them. This holds immense potential for complying with recent AI regulations, which require automated decision-making processes to be supervised by humans to mitigate the risks of discrimination. By only having to review instances at high risk of unfairness, our framework can make this process more practical and time-efficient. To further empower human users, further research could involve human experts in the selection of *at-risk* subgroups and in choosing distance function and parameters for Situation Testing. Also, user studies can help in understanding how humans engage with such a system. For this, one should consider adding explanations for all non-rejected instances, so that humans can still explore the base classifier in the accepted cases.

Chapter 4

Introducing A Benchmark Dataset

The previous chapters of this thesis have dealt with tools and algorithms for bias detection and mitigation. The reoccurring theme and motivation behind each chapter was that traditional approaches that measure or optimise fairness through single numeric metrics are insufficient for thoroughly understanding and resolving discrimination. Instead of assessing the fairness of one entire system through one number, efforts should be made to understand and fix unfairness exactly where it occurs. This chapter further builds on this idea, this time exploring it from the perspective of evaluating the effectiveness of fairness interventions. We have collected a dataset with full information about which instances are affected by discrimination. To do so we started with an already existing dataset with information about high school students and their leisure- and study behaviour. We assumed that the current version of the decision labels, showing whether students passed or failed some exam, is fair and collected a biased version of these labels through a human experiment. In this experiment we captured realistic human stereotypes by letting participants predict students' study performance based on the characteristics available about them. We show how this new version of the labels is biased against boys, and illustrate how the effectiveness of fairness interventions can be evaluated on the data by applying them on the biased version and testing them on the fair one. Further, we highlight the shortcomings of the traditional evaluation scheme given by the fairness-accuracy trade-off. We show how some interventions that perform well according to this trade-off do not necessarily perform well with respect to the unbiased labels in our dataset[1] [2]*.*

[1] This chapter is based on: *Lenders, D., & Calders, T. (2023, March). Real-life performance of fairness interventions-introducing a new benchmarking dataset for fair ML. In Proceedings of the 38th ACM/SIGAPP symposium on applied computing (pp. 350-357).*

[2] A "Datasheet" to give a detailed yet concise description of our new dataset is available in the Appendix B.1

4.1 Introduction

As we have highlighted in the introductory chapter of our thesis, over the years many methods have been designed to mitigate bias in ADM systems and improve the fairness of a decision-making process. As we also have highlighted in this chapter, evaluating the effectiveness of such fairness interventions is far from arbitrary. To reiterate that point, take the decision task visualized in Table 4.1 with information about a loan applicant: a banker decided not to grant her a loan, and Model A agrees with this decision, while Model B disagrees. We can easily evaluate the accuracy of the models, but it is much harder to evaluate their fairness. After all, the negative label might have been the result of the banker being biased against women. If we knew that the applicant would have paid back the loan if she had received one, we can say that our data set contains label bias and that Model B is fairer than Model A, because it corrected this bias. When adopting this point of view, we also see that, while Model A is more accurate in regards to the label assigned by the banker, Model B is more accurate in regards to the deserved label.

Though assuming the existence of a "fair" and "biased" version of the label can theoretically help in evaluating fairness interventions, it is hard to apply in practice, as we usually do not know which instances are affected by label bias. Therefore, researchers typically evaluate their interventions by examining the accuracy-fairness tradeoff on the labels at hand. They deem an intervention as successful if it satisfies a fairness definition of choice while sacrificing little predictive performance on these labels [54, 73, 89]. It is easy to see how this evaluation scheme is not optimal, as it makes little sense to strive for high accuracy on labels that are not believed to be “true" in the first place.

Table 4.1: Without access to an unbiased label, it is difficult to evaluate the fairness of Model A and B

Sex	Credit Amount	Job Status	....	Bank Decision	Model A	Model B	Deserved Decision
Female	10k	Employed	....	No Loan	No Loan	Loan	?

Because assuming the existence of “fair" and “biased" labels overcomes the shortcomings of this traditional evaluation scheme, some researchers use simulated data to test their algorithms. One example of this is in Chapter 2 of this thesis, where we adopted a fully simulated approach, characterizing the complete joint distribution of a sensitive attribute, all legally grounded attributes, a biased, and an unbiased version of the decision label with a Bayesian Network. A similar approach was taken by [137]. Other researchers use semi-simulated approaches where they start from an existing dataset, apply some operations to make it bias-free, and then randomly add bias to it again, to have full information about which instances are affected by this bias [51, 137, 144]. They then evaluate their algorithms, by measuring how well they can predict the fair labels after being trained on the biased ones. Though the advantage of these controlled settings is obvious, there is no denying that these approaches cannot capture the complexity of realistic data and its biases. For instance, in previous studies, bias was often introduced by arbitrarily altering the decision labels of individuals from protected groups without considering the specific characteristics of those individuals [51, 137, 144]. In reality, discrimination is influenced not only by protected group membership but also by other personal attributes. Hence, randomly changing labels fails to reflect the intricate dynamics of discrimination. Additionally, in many experiments researchers hold the simplistic "We’re all equal" assumption: in a fair world, a sensitive attribute is not correlated with a person’s other characteristics and their eligibility for a positive decision outcome [137, 144]. Again, this assumption is unlikely to reflect real-life, where a person's sensitive information (like their age or sex) may very well be correlated to their job, education, or other factors relevant to a decision problem.

Another final disadvantage of tests on synthetic data is the risk of developers "making up the data to suit the algorithm" [47]. When developers develop a fairness intervention, they are likely to have some assumptions about the dynamics behind discrimination and the way it has emerged in the data. If they base their simulated data on the same assumptions, tests on this data will give an unrealistically optimistic view of the performance of the intervention.

In this chapter, we address the problem of evaluating fairness interventions without synthetic data by introducing a new dataset[3]. This data consists of real-life information about students, their free time and study behaviour. As a fair version of the decision labels, the dataset contains information about whether students passed a course, while we obtained the biased version of the labels through a human experiment, asking participants to estimate the students' performance based on information about their demographics and personality. We show how the latter version of the decision labels is biased against male students, and explore some interesting discriminatory patterns in the data. We proceed with describing the results of a small benchmarking study, to demonstrate how our dataset can be used to evaluate fairness interventions and how it leads to new insights about their performance, that would not be gained using traditional evaluation schemes. Before concluding this chapter, we describe other use cases for our dataset, with which we hope to encourage future research.

4.2 Relation to prior work

New Datasets for fair ML The dataset we introduce in this chapter is a contribution to the work on better datasets for algorithmic fairness. This line of work has emerged from the criticism towards the datasets currently used by the fair ML community, regarding their data quality and the relevance of their associated prediction tasks [9,47]. Efforts have been put into introducing new benchmarking datasets [17, 40, 82], with "folktables" by Ding et al. being one prominent example [40]. Though we recognize these contributions, we note that these datasets do not facilitate the objective evaluation of fair ML algorithms as they do not provide both a fair and biased version of their decision labels. To the best of our knowledge, our dataset is the first realistic one doing so.

ML for discrimination prevention Our dataset is intended for testing fair ML algorithms that assume that the training data they are based on contains discrimination. Mehrabi et al. [93] define "discrimination" as a type of bias, that arises when human decision-makers have prejudices against specific groups of people and give them different decision labels than they deserve. In other publications, this type of bias is also referred to as label bias [31, 70, 137]. Typical examples in which discrimination/label

[3] The collected dataset is under license CC BY-SA 4.0 as csv file available online: `https://www.kaggle.com/datasets/daphnelenders/performance-vs-predicted-performance`

bias occur are hiring processes or loan applications. Note how discrimination is different from other types of data bias like representation bias, where the sampling procedure of the data is biased but the data labels themselves are assumed to be correct [93]. There are many fair ML algorithms specifically made to mitigate discrimination (as a special type of data bias); some examples include [70, 89, 143, 144]. Developers of such algorithms can greatly benefit from our new dataset, by testing how well their methods can detect/remove the discrimination present in the biased version of our labels to obtain the fair ones. Note, that our data may still contain unfair patterns in terms of the opportunities students have when preparing for an exam (e.g., family support, being able to afford tutors). Our dataset should not be used to see if ML algorithms can make up for these kind of inequalities.

4.3 Creating a Dataset with a Biased And Fair Label

To create a realistic dataset with a fair and unfair version of the decision labels, we based our data collection on an already existing dataset, that is publicly available online. It is called the “Student Alcohol Consumption" dataset and consists of entries of high school students following a course [34]. For each student, the grade is recorded for three exams, as well as some demographic information (e.g., their sex or age) and information about their free time behaviour (e.g., how often they go out) and study behaviour (e.g., their study time). The decision task we are interested in is to predict whether students have a passing grade for the third exam of the course, whereas grades are measured on a scale from 0 to 20 and the lowest passing grade is 10. We started from the assumption that the current version of the labels in the data is fair in regard to sex of the students. This assumption is supported by the following observations:

- Every student had the chance to write the exam and prove their capabilities in the subject. This setup is different than in other decision tasks, like loan applications, where individuals that are denied a loan, do not get the chance to prove whether they would have deserved it;
- The grades of an exam can be somewhat objectively measured since teachers typically make use of pre-defined rubrics when doing so;
- The positive decision ratio for girls in this dataset is 84,34% and with that only 3,5% higher than for boys. Thus, at least on a group level, the distribution of passing grades between boys and girls seems to be fair.

To obtain the biased version of these labels we conducted a human experiment, where humans had to predict students’ exam performance based on limited information about them. Because participants only saw this limited information and not the written exams, we expected them to rely on stereotypes when making the predictions [79]. In particular, we expected them to be biased against male students, as there are many stereotypes about boys being less mature and more lazy throughout high school [19]. Further, we expected that the introduced bias would not only be based on students’ sex but would also interact with their other characteristics, like their free time or study behaviour. In other words, we expected their bias to be complex and messy, like in real life.

To make sure that these expectations were met, we first conducted a proof-of-concept study, where we confirmed that humans indeed have stereotypes against boys when predicting their school performance. With the same experiment, we found that in cases where inherent biases are not present, they can be triggered through stereotype activation. This is a process in which stereotypical information is presented, and which can lead people to apply the presented stereotypes in decision processes [136]. In the next section, we give a short description of this preliminary study and describe how we utilized the results to conduct our main study. In section 4.4 we proceed with a description of the resulting dataset and in section 4.5 we provide a case study to show how it can be used to evaluate fairness interventions.

4.3.1 Proof-of-Concept Study

Our study design for our main experiment was based on a proof-of-concept study[4]. In the task of this study, different from the main study, we presented the same eight student profiles to the participants, containing basic information about each student, for which participants had to make grade predictions. The main manipulation was that part of the participants was presented with a version of these profiles for which the sex was swapped. In other words, a profile that belongs to a female student was, in this condition, said to belong to a male student (and vice versa). By comparing the predicted grades across these two conditions, we found that for some students, participants predicted lower grades for the male version of the profile than for the female version. This showed that, as expected, participants have an inherent bias against boys when making grade predictions. Since we only found this effect on some student profiles, this also confirmed our hypothesis that the bias introduced in our experiment does not only depend on students' sex but also their other more complex characteristics.

The second goal of our proof-of-concept study was to see if bias against boys can be triggered. For this purpose, our second experimental manipulation was exposing participants to some form of stereotype activation before the prediction task. Next to a baseline condition (where no stereotype activation was presented), we included two stereotype activation conditions, in which information was presented that suggested that boys perform less well in high school than girls (for a more detailed description see section 4.3.3). By comparing the difference in grades assigned to male and female versions of the profile across different stereotype activation conditions, we found that in some cases discrimination can indeed be triggered through stereotype activation. In other words, when stereotypical information about boys is presented, participants are more likely to assign a lower grade to a male than a female version of a student profile. We found this triggered bias an interesting addition to the already present biases against boys and therefore decided to include the different stereotype activation conditions also in our main study.

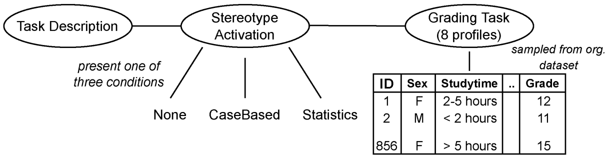
In general, the results of our proof-of-concept study showed that the used study setup is appropriate to elicit interesting biases in human decision-makers. We, therefore, decided to keep a similar study setup (regarding the materials that were used and the nature of the grading task) for our main study, where we collected a biased version of the decision label for all students of the original data. We will describe these materials and the exact task in more detail in the next sections.

[4] More information on this study can be found in the Appendix B

4.3.2 Main Study

The aim of our study was to collect a biased version of the decision labels for all students in the “Student Alcohol Consumption" dataset. For this, we used the study design as illustrated in Figure 4.1. As in our proof-of-concept study, the main task for the participants was to make grade predictions for eight student profiles. However, this time we did not present the same eight profiles to each participant, but instead randomly sampled ones from the original dataset until for every instance we had one biased grade prediction.

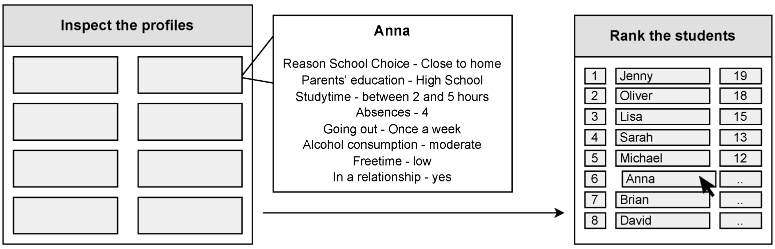
Before making the grade predictions, participants were (just like in our proof-of-concept study) exposed to one of three stereotype activation conditions. These conditions, along with the rest of the material for the study are described in the next section[5].


Task Description
Stereotype Activation
Grading Task (8 profiles)
present one of three conditions
None
CaseBased
Statistics
sampled from org. dataset

ID	Sex	Studytime	..	Grade
1	F	2-5 hours		12
2	M	< 2 hours		11
856	F	> 5 hours		15

Figure 4.1: Experimental setup to collect a biased label for the students of the “Student Alcohol Consumption" dataset

4.3.3 Grading Task & Materials



Anna

Reason School Choice - Close to home
Parents' education - High School
Studytime - between 2 and 5 hours
Absences - 4
Going out - Once a week
Alcohol consumption - moderate
Freetime - low
In a relationship - yes

1	Jenny	19
2	Oliver	18
3	Lisa	15
4	Sarah	13
5	Michael	12
6	Anna	..
7	Brian	..
8	David	..

Figure 4.2: Illustration of the grading task. Participants are presented with eight student-profiles, which they have to rank according to their expected exam performance. In this interface they are also prompted to enter a grade prediction

In Figure 4.2 we give an illustration of the task setup of our study. In the main task, each participant was presented with eight student profiles, half of which male and the other half female, extracted from the “Student Alcohol Consumption" dataset. To be able to present an equal number of male and female profiles to each participant, we randomly excluded 133 female instances from the original dataset, through which we ended up with a total of 856 instances. By making sure that half of these were male and half female, we ensured that our final dataset does not contain any representation bias in regards to the “sex" of the students, but instead only the label bias as introduced by our experiment.

[5] Our study was approved by the Ethics Committee of the University of Antwerp, under reference number SHW_21_128.

While the original dataset contains more than 30 attributes to describe each student, we chose to only present eight of them per profile, as to not overload participants with information. As attributes, we chose those that had high variability between students and that could in legitimate or stereotypical ways be associated with school performance. In Figure 4.2 we show an example of such a profile.

Just like in the figure, each student profile was presented in a tabular format. To convey the sex of each student, we randomly assigned each profile to one of four male/female names, depending on the students' sex in the original dataset. These names were chosen to represent common names in English-speaking countries. In the experiment, all eight profiles were presented on one page, where the order of presentation was randomized. On top of this page, participants were presented with a list of all student names followed by a blank field. They were asked to use a drag-and-drop interface to rank the students according to their expected performance. Additionally, they were prompted to enter specific grade predictions (ranging from 0 to 20) in the blank field next to each student's name.

Before the grading task, participants were exposed to one of three forms of stereotype activation:

1. **None** - Baseline condition in which no extra information is presented.
2. **CaseBased** - Here we presented participants with three student profiles along with the grades of the students. Two profiles belong to male students with low grades (5/20 and 10/20), while one belongs to a female student with a high grade (17/20).
3. **Statistics** - Here we presented a graph showing statistics about how some risk factors affect boys' chance to pass an exam more than they affect girls' passing chances. One presented risk factor was, e.g., having more than 6 school absences, which makes boys ~15% more likely to fail, while girls only ~4% more likely.

As mentioned before, we learned from our proof-of-concept study that stereotype activation conditions can negatively influence participants' grade predictions for male students. Hence, these conditions were added to our main experiment, to add another layer of bias to the already inherent bias against boys. We also considered the different conditions as a reflection of real life, where different decision-makers may be exposed to a different set of assumptions about their decision subjects.

4.3.4 Participants

We recruited our participants through social media channels and the survey exchange platforms SurveySwap and SurveyCircle. To participate, a consent form needed to be filled out. We continued the data collection process until for every student in the original dataset we had one participant making a grade prediction for them. Full information about the participants of our experiment is included in Appendix B.

4.4 Resulting Dataset

In this section, we are going to explore the grade predictions for the students of the “Student Alcohol Consumption" dataset, in particular how they are biased against male students. Starting from there we show how these grade predictions can be converted to binary labels, and how they then serve as a biased version of the already present labels of the dataset.

4.4.1 Bias in Grade Predictions

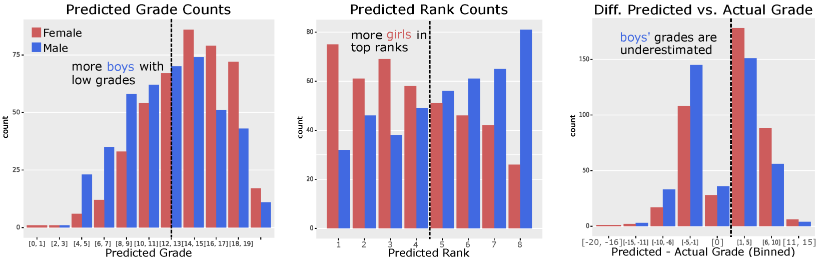


Figure 4.3: The performance predictions in our dataset are biased against male students

To get an idea of whether our study participants had biases based on students’ sex, we first look at Figure 4.3. Here we visualize the distribution of predicted grades per sex, the difference between students’ actual grades and their predicted grades, and the ranking positions they were assigned to (since each participant ranked and graded 8 students, these positions range from 1 to 8).

Judging from the figures it appears that, just like in our proof-of-concept study, participants were biased against boys in their grade predictions. Looking at the left figure, we see that lower grades (grades lower than 13) are more frequently assigned to boys, while higher grades are more frequently assigned to girls. The same bias is reflected in the distribution of ranking positions across the sexes: girls are more represented in the top ranks (between 1 – 4), and boys in the lower ranks (rank 5 – 8). Lastly, we see in the right figure, that this predicted performance difference is not reflected in the actual grades of students: when looking at the differences between both, we see that boys’ grades were often underestimated, while girls’ grades were often overestimated.

Though the predicted grades and ranks are already interesting to observe, they still have to be translated into binary labels, to make them useful for the typical fair ML benchmark setting.

4.4.2 From Grade Predictions to Binary Labels

The most obvious way in which to translate the grade predictions to binary labels is to check whether the students would pass or fail the exam according to the predicted grades (where grades ≥ 10 are passing grades). In Table 4.2a and 4.2b we show how then the biased labels relate to the actual labels, for boys and girls separately.

Table 4.2: Relation between biased and fair labels as obtained through different conversion strategies

(a) Girls *(pass-fail strategy)*

Actual Predicted	Pass	Fail
Pass	319 (74.54%)	56 (13.08%)
Fail	42 (9.81%)	11 (2.57%)

(b) Boys *(pass-fail strategy)*

Actual Predicted	Pass	Fail
Pass	267 (62.38%)	44 (10.28%)
Fail	79 (18.46%)	38 (8.88%)

(c) Girls *(ranking strategy)*

Actual Predicted	Pass	Fail
Pass	307 (71.73%)	17 (3.97%)
Fail	54 (12.62%)	50 (11.68 %)

(d) Boys *(ranking strategy)*

Actual Predicted	Pass	Fail
Pass	241 (56.31%)	7 (1.64%)
Fail	105 (24.53%)	75 (17.52%)

Another possibility is to use the ranking position assigned to each student. One could, e.g., change the passing label of all lowest two ranked individuals to “False", and change it to “True" for the highest two ranked individuals (results shown in Table 4.2c and 4.2d).

With both conversion strategies, we see that female student benefit from a higher true positive rate (0.8837 and 0.8504 respectively for the pass-fail and the ranking strategy) than boys (0.7946 and 0.6965), while the girls’ biased labels also contain more false positives (0.8358 and 0.2537) than the boys’ (0.5366 and 0.0854). In other words, girls are more frequently predicted to pass the exam when they would actually fail, while boys are more frequently predicted to fail the exam when they would actually pass. This shows, that the collected data can be used in testing fair ML algorithms, that treat male students as the disadvantaged group. Further, this data gives a more complex and interesting perspective on bias than simulated data, where usually only instances from the disadvantaged group do not get the decision label they deserve.

For the experiments in the remainder of our chapter, we are going to use the binary labels as obtained by the ranking strategy. This choice was based on the fact that the labels obtained by the pass-fail strategy contain a lot of false positive individuals, while most fairness interventions focus on false negatives (i.e. individuals who were assigned a negative label when they deserve a positive one).

4.4.3 Subgroups affected by bias

Table 4.3: Subgroups with highest false negative (left) and false positive rates (right)

(a) Subgroups with highest false negatives rates (b) Subgroups with highest false positive rates

Subgroup	**Size**	**#FN**	**Subgroup**	**Size**	**#FP**
studytime == ’less than 2 hours’	272	94	alcohol consumption == ’low’ AND sex == ’F’	189	11
romantic rel. == ’no’ AND ’studytime’ == ’less than 2 hours’	196	75	freetime == ’average’ AND goout == ’Thrice a week’	90	8
sex == ’M’ AND studytime == ’less than 2 hours’	184	66	sex == ’F’	428	17
alcohol consumption == ’very high’	191	66	alcohol consumption == ’low’ AND sex == ’F’ AND romantic rel. == ’no’	112	8
romantic rel. == ’no’ AND sex == ’M’ AND studytime == ’less than 2 hours’	142	56	romantic rel. == ’no’ AND sex == ’F’	264	12

We are now going to look more closely at the biases introduced by our experiment and see which subgroups are most affected by high false positive and false negative rates

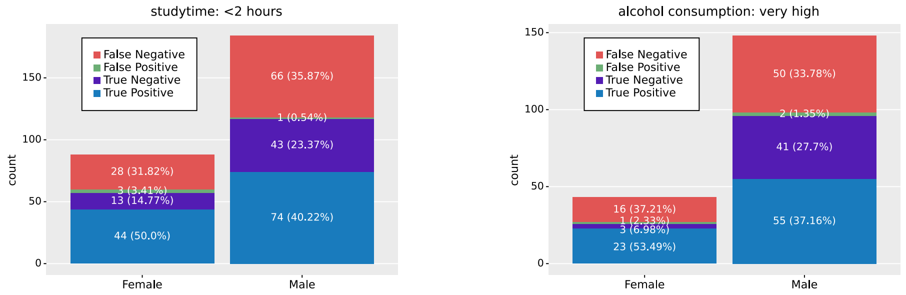


Figure 4.4: Two subgroups suffering from high false negative rates (as detected by the apriori algorithm). Even though both subgroups are not explicitly male/female we see that in both groups there is a majority of male students, explaining higher false negative rates for boys in the overall data

in their biased labels. For this purpose, we applied the apriori subgroup discovery, as developed by [77], on our data. This algorithm finds the most interesting associations between variables in the data and some property of interest (in this case being a false positive or false negative), as measured by some quality measure of choice. In our case we used the “weighted relative accuracy gain", a measure that both takes the strength of the association between a subgroup and the target into account, as well as the size of the subgroup (where bigger subgroups are seen as more interesting). In Table 4.3 we see the top 5 subgroups with highest false negative and false positive rates.

One thing that immediately becomes apparent is that many of the false negative subgroups are explicitly male, while all top 5 false positive subgroups are female. Further, we observe in both clusters of subgroups some stereotypes of typically “good" and “bad" students. For instance, students with high alcohol consumption and low amount of studytime are likely to be predicted to fail (when they actually pass), while students with high studytime and low alcohol consumption are likely to be predicted to pass (when they actually fail). Further, there are some student characteristics, which are in more surprising ways connected with the bias patterns: in particular, we see that not being in a romantic relationship is associated with both being a false negative and being a false positive. Also, we see that going out regularly (thrice a week) is associated with being a false positive, which may be surprising given that this trait may stereotypically be associated with low-performing students. Both unexpected findings may however be explained by the fact that there are not many false positives in the biased labels and that it may be hard to find strong patterns of them.

We observe another interesting finding when inspecting the false negative subgroups that do not explicitly contain the ’sex’ of the students, like the subgroup of students with a short studytime and the subgroup of students with a very high alcohol consumption. In Figures 4.4 we observe the distribution of male and female students over both subgroups, as well as how their biased label relates to their actual one. Already at first glance, we see that both subgroups contain a much higher proportion of male than female students. Relatively speaking, there is no big difference in the proportion of false negative instances within these subgroups; in the subgroup of students with high alcohol consumption, there are even higher false negative proportions for girls (37.21%) than for boys (33.78%). However, because there are that many more male students within both subgroups, absolutely speaking a larger number of boys are affected by the false negatives, which then leads to higher overall false negative proportions among boys when looking at the data as a whole. The combination of stereotypical biases and more complex patterns like these, further highlights the overall appeal of our dataset, especially as an alternative to simulated data.

4.5 Use case - Testing fair ML algorithms

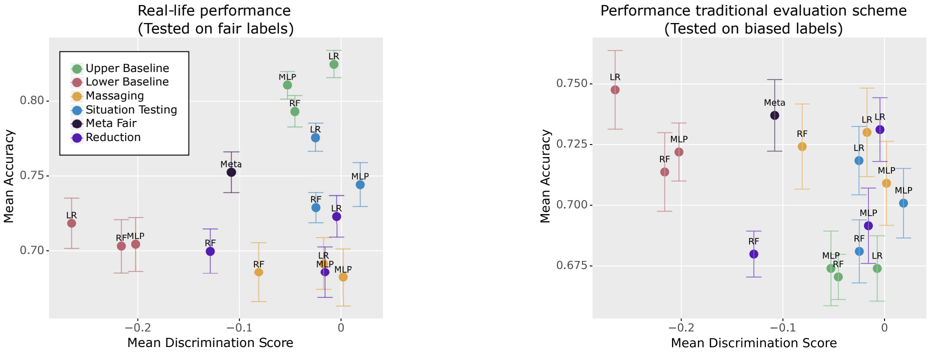


Figure 4.5: Experiments on our new dataset reveal that some interventions that perform well according to the traditional evaluation scheme (i.e. good accuracy-fairness tradeoff on biased labels), do not perform well in regards to the fair ones

In this section, we illustrate how our dataset can be used to test the effectiveness of fairness interventions. First, we test two pre-processing interventions: the “massaging" approach [73] and the “situation testing" algorithm [89]. With both interventions, label bias is detected in the data, such that it can be removed and a classifier can be trained on a de-biased version of it. Further, we test two in-processing approaches, namely “meta fair learning" [25] and “exponentiated gradient reduction" [1]. Both algorithms aim to mitigate discrimination at the training stage. The meta-fair algorithm does so by building a classifier that maximizes accuracy under a given fairness constraint. The reduction algorithm takes a base learner (i.e. any standard classifier) and reduces its learning algorithm to a cost-sensitive classification problem, where iteratively different instances are assigned different error weights until a fairness goal is achieved (while simultaneously minimizing the classification error). For both in-training techniques, we set “demographic parity" as the fairness goal, with which we aim to obtain equal positive decision ratio’s between boys and girls.

We evaluate all interventions by how well the resulting classifiers can predict the fair labels of a held-out test set, after being applied on the biased labels of the training set. In this case, we used the biased version of the labels obtained by the “ranking" strategy (see the previous section). Next to the four fairness interventions, we included an upper and lower baseline: the upper baseline is a classifier trained on the fair version of the labels and the lower baseline is a classifier trained on the unfair version without applying any intervention.

We tested the performance of our baseline methods, the pre-processing algorithms and the “exponentiated gradient reduction" algorithm in combination with three classifiers:

a Multi Layer Perceptron, a Random Forest and a Logistic Regression classifier. As the meta-algorithm does not rely on any external classifier and instead builds its own classification model, we only report the results of this model itself. We used 10-fold cross-validation and report the average accuracy over all 10 test sets, as well as the average discrimination score. The discrimination score is measured as the difference in positive decision ratios between boys and girls. Since in the fair labels of the dataset, nearly the same ratio of boys and girls passes the exam, a fair classifier should have a discrimination score of approximately zero. Further, since we are testing each intervention on the fair labels of the dataset, the predictive accuracy should be as high as possible.

We also include results of the same experimental setup, where we evaluate each intervention on the biased labels. This corresponds to the traditional evaluation scheme of fair ML, where we have no access to the fair labels and aim for the best accuracy-fairness tradeoff on the biased ones. The results are given in Figure 4.5[6].

When assessing the performance on the fair labels, we see that both in regard to accuracy and fairness, the upper baseline performs best while the lower baseline performs worst. The results of the fairness interventions are more surprising. We see that when applying "massaging" or the "reduction" technique, the discrimination score of the predicted labels has improved in comparison to applying no intervention. However, regarding the accuracy, it performs worse. This clearly indicates that these interventions lead to some form of window dressing: they ensure that a more equal ratio of boys and girls get a positive decision outcome. However, they do not guarantee that the people assigned to a positive label also deserve one. The results of the situation testing approach seem more promising, especially using a logistic regression classifier after applying this method yields satisfactory results. However, as with the "exponentiated gradient reduction" method, it appears that the effectiveness of the intervention can depend a lot on the classifier it is combined with. Finally, we see that with the "meta-learning" technique we obtain an acceptable accuracy but also a relatively high discrimination score, indicating that this intervention is not appropriate for this specific decision task. Note, that the performances on the biased version of the labels, sketch a quite different picture. If we would evaluate interventions based on their accuracy-fairness trade-off on the biased labels, we would deem "massaging" or the "reduction" technique as effective techniques, not recognizing how these interventions degrade the accuracy on the fair labels.

While it lies not in the scope of this chapter to provide a deep analysis of why some fairness interventions are more successful than others, the results show how our benchmark dataset brings new insights about fair ML algorithms, that would not be found using traditional evaluation approaches. Developers of fair algorithms could apply a more detailed error analysis to understand why/where their interventions fail. In particular, it could be interesting to see whether certain interventions work well for some subgroups, but have blind spots for others and how to improve performance on the latter. Further, our dataset could be used to understand how the effectiveness of the fairness intervention may depend on its hyperparameters[7] or the classification models they are combined with.

[6] For full implementation, see: `https://github.com/calathea21/benchmark_data_analysis`

[7] we used the default ones of the AIF360 library: `https://aif360.readthedocs.io/en/stable/`

4.6 Suggestion for other Use Cases

Our dataset can be used to evaluate the effectiveness of fair ML interventions, but there are also other interesting use cases for it. In this section, we are going to highlight three of them.

Exploiting knowledge about stereotype activation When collecting our dataset, participants were exposed to different kinds of stereotype activation. It would be interesting to see whether exploiting the information about which profile was graded under which stereotype activation condition, can improve the effectiveness of a fairness intervention. In real-life, this may translate to situations in which we know about the different circumstances in which decision-makers operate. If we, e.g., know that half of the bankers deciding on loan applications received a different training than the other half, this may help in removing the different sorts of biases that may have been introduced by both groups.

Using portion of fair labels for transfer learning While it is not possible to access a complete version of the fair decision labels for each dataset, there are situations in which a small portion of the fair labels is available: for instance in the context of loan applications, we may not know for the persons who were denied a loan whether they would have deserved one, but we do know for the individuals that received one, whether they paid it back. Using this small portion of the "fair labels", it might be possible to train a classifier to then make predictions for all instances for which we do not know the fair labels. Our dataset could be used to study the general viability of such an approach.

Incorporating domain knowledge Whether students get discriminated does not only rely on their sex, but also on other characteristics that stereotypically are connected with low school performance. This suggests that it may be beneficial to involve a domain expert, with a deep understanding of the dynamics behind such stereotypes, in a fairness intervention. With our dataset, it is easy to evaluate which ways of incorporating domain knowledge are successful, and the results may generalize to other decision tasks.

4.7 Limitations & Conclusion

In this chapter we showed how we created a dataset, that can be used to benchmark fair ML algorithms. We based this data on an already existing dataset, which contains information about students and whether they pass an exam or not. We assumed that the current version of this decision label is fair and collected a version of these labels, which are biased against boys, through a human experiment. With this new dataset, we facilitate the evaluation of fair ML interventions, by seeing how well they can predict the fair labels after being trained on the biased ones. Our data overcomes the shortcomings of simulated alternatives since it is more complex and realistic in terms of its introduced bias.

While our dataset has many potentials for the fair ML community, users should be aware of its limitations. First, in the collection of our dataset we treated the “sex" of the students as a binary variable. This is a direct consequence of the fact that no information on self-reported (non-binary) gender identity was available in the original “Student Alcohol Consumption" dataset. We acknowledge that this limits the validity of our study results as well as our final dataset, as people who do not identify with a binary gender category may face very different levels of discrimination, that should be accounted for in real-life as well as algorithmic decision processes.

Further, the dataset is with a total of 856 instances, rather small. Many real-life datasets for which fairness concerns arise are much bigger. Hence, at least in this regard, our data may not capture the full complexity of real-life applications. Relating to this, researchers should be cautious not to overgeneralize the results from experiments on our data. While this goes with any type of benchmarking data, it is especially important in the context of fairness. Whether a fairness intervention can be considered effective is dependent on the domain of a decision task and the data at hand. Still, given the full information about label bias in our data, using it as one test case is worthwhile to get a high-level idea of an algorithm's performance and its blind spots.

Lastly, we highlight that our data was made to test the effectiveness of fair ML interventions that target discrimination. As previously mentioned, before collecting the biased version of our labels, we assumed that the original version of the labels is fair. While this assumption is true in the sense that every student had the opportunity to prove their capabilities on an exam, it may not be true in regards to whether everyone had the same opportunity when preparing for it (e.g., some students might have been able to afford private lessons). If researchers are interested in seeing whether their fairness interventions can compensate for such inequalities, our dataset should not serve as a benchmark.

Chapter 5

Outlook

The previous chapters of this thesis have zoomed into various aspects regarding fairness in automated decision-making algorithms, focussing on specific tools and algorithms for bias detection, mitigation and the evaluation of their effectiveness. The following chapter will zoom out from these specific considerations and give a broader overview of the research surrounding algorithmic fairness, how it has developed over the years and where its biggest gaps lay. For this purpose we present a scoping review of the literature over the past fifteen years, utilising sources from Web of Science, HEIN Online, FAccT and AIES proceedings. All articles come from the computer science and legal field and focus on AI algorithms with potential discriminatory effects on population groups. We annotated each article based on their discussed technology, demographic focus, application domain and geographical context[1] *and analysed the evolution of the literature regarding these characteristics. Though we observe a growing trend of literature addressing a broader variety of topics and becoming more specific, a substantial portion of contributions remain generic and only discuss algorithmic discrimination in the context of classification systems without concentrating on the domains these systems operate in or the demographic groups they harm. Regarding the geographical context of research, the focus is overwhelming on North America and Europe (Global North Countries), with limited representation from other regions. This raises concerns about overlooking other types of AI applications, their adverse effects on different population groups, and the cultural considerations necessary for addressing these problems. With the help of some highlighted works, we advocate why a wider range of topics must be discussed and why domain-, technological, diverse geographical and demographic-specific approaches are needed. This chapter also explores the interdisciplinary nature of algorithmic fairness research in law and computer science to gain insight into how researchers from these fields approach the topic independently or in collaboration. By examining this, we can better understand the unique contributions that both disciplines can bring to move the research field forward.*[2]

[1] The data is available at `https://github.com/calathea21/algorithmic_fairness_scoping_review`

[2] This chapter is based on the following paper: *Lenders, D. & Oloo A. (Under submission). 15 Years of Algorithmic Fairness: Scoping Review of Interdisciplinary Developments in the Field.* Daphne Lenders is first author of this paper. Both authors contributed equally in developing the research idea and questions, but Daphne was main responsible for the quantitative result analysis

5.1 Introduction

Research on algorithmic fairness has been present for about 15 years. What initially started as a slow movement has become a popular and prominent research field, with dedicated conferences about the topic, like ACM FAccT and AAAI AIES. Throughout these years, the field has kept evolving, fueled by public discourses about unfair algorithms, new legislations around AI and ever-emerging technologies. While it is generally well known that the field develops rapidly, less is understood about how it has developed, what the most prominent research areas are and where the research efforts come from. Yet, only when zooming out and having a better view of the large body of literature that already exists, we get an idea of whether the research has kept up with the pace in of technology and where the biggest research research gaps and opportunities lay. For this purpose, we have conducted a scoping review on the field of algorithmic fairness. Using four scientific databases, namely, Web of Science, Hein Online, ACM FAccT and AAAI AIES proceedings, we have sampled a total of 1570 papers dealing with this topic and have annotated them in terms of the domain they consider, the demographic groups they focus on and technology they discuss. By providing aggregated results over these three metrics, we sketch an overview of the most prominent research areas within the field, and how these have developed over the years. In doing so, we also differentiate between the research efforts coming from primarily Computer Science and Law based perspectives. We highlight how authors with different expertise approach research areas differently, and which areas remain under-addressed by either or both communities. By then highlighting some research studies in less popular areas of the field, we emphasize which areas need to be addressed to tackle algorithmic discrimination in all of its forms, rather than limited to a narrow set of technologies and domains. To summarize, the first part of our work addresses the following research questions:

RQ1: How has algorithmic fairness literature developed in terms of the domains they address and what are the opportunities/gaps in adopting domain-specific approaches from a technological and legal perspective?

RQ2: How has algorithmic fairness literature developed in terms of the demographic groups they focus on? How does this differ between researchers with technological and legal expertise?

RQ3: How has algorithmic fairness literature developed in terms of the technologies they address? How does this differ between researchers with technological and legal expertise?

Our last research concerns the geographical context of the research on algorithmic fairness, both in terms of the authors' affiliations and the geographical areas they address. We showcase how much of the current literature is primarily centred around Global Northern countries and highlight how more recent contributions, focussing on other geographical areas, bring to light important considerations around algorithmic fairness that should not be overlooked. Hence the last research question of this study is:

RQ4: What is the geographical scope of algorithmic fairness literature, both in terms of researchers' geographical affiliation and the content of their papers?

5.2 Related Literature & Motivation

There are many literature reviews available related to algorithmic fairness. Different from scoping reviews, these works dive into specific aspects of the topic, like bias mitigation methods for classification algorithms [66], datasets commonly used in experiments [47], or fairness concerns related to specific technologies like computer vision system. [91]. Their goal is to summarize the most important contributions and insights surrounding these topics and identify concrete research gaps related to them. In comparison, scoping reviews on algorithmic fairness are much more sparse. Rather than summarizing the literature on one concrete topic, scoping reviews aim to give a high-level overview of broad and general research areas that encompass many different technologies, domains and disciplines. Scoping revies aim to sketch the breadth of these areas and identify the most popular research directions. In doing so, they also highlight which areas are currently underexplored and need more attention from the research community.

Vilaza et al. (2022) report a scoping review on ethics in technology and inspect 129 papers coming from the SIGCHI conferences. In particular, they assess the themes of the ethical considerations in each paper (e.g. privacy, discrimination, mental well-being etc.), the population groups that are discussed, and the type of technologies inspected (e.g. web applications, social media, etc.). Similarly, a study by Birhane et al. (2022) dives into the topic of AI ethics across FAccT and AIES papers. They aggregate results of 535 papers, focusing on how concrete or abstract each work of literature is regarding the ethical aspects they address. In particular, they inspect whether papers discuss case studies of algorithmic systems already used by industry, and how much effort the works put into understanding how real stakeholders are affected by these systems. A study that emphasizes geographical regions/contexts in which AI ethics are addressed is conducted by Urman et al (2024). Specifically, they inspect 200 papers describing AI auditing studies, not just identifying which ethical aspects the AI systems are audited for, but also highlighting the countries on which the audits were focused, and the geographical affiliation of the authors contributing to these studies. Our contribution sets itself apart from these already existing scoping reviews in various ways:

1. Different from other studies, we focus on algorithmic fairness as one sub-area of AI ethics, rather than AI ethics in general. This allows us to identify the research landscape and gaps more specific to this area, addressing the research focus in terms of addressed domains, demographic groups, technologies and geographical context

2. We are the first scoping review, to inspect the development of the research area from an interdisciplinary perspective, focusing on how authors with Computer Science and Law expertise address this topic differently, and where the research gaps in either or both of the fields lay

3. To the best of our knowledge our study is the largest scoping review on AI ethics, aggregating the results of a total of 1570 papers. By not merely focussing on contributions coming from FAccT and AIES, we get a better overview of the current literature.

5.3 Methodology

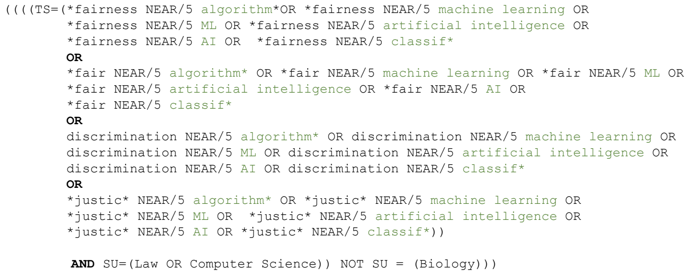
To conduct our scoping review we adopt the PRISMA (short for: “Preferred Reporting Items for Systematic Reviews and Meta-Analyses") guidelines [88]. This means that our methodology consisted of three key steps: the first, was devising a search strategy, by selecting the scientific databases for locating relevant papers and designing queries to search these databases. The second step was going through the found papers and deciding which ones to include in this review. The third and last step was annotating the selected papers for relevant information and analysing the results. We are going to describe each step in more detail in the following sections.

5.3.1 Databases & Search Query

We used Web of Science as our main database for scientific articles. Using their advanced search function, we set up the search query as seen in Figure 5.1 to find papers related to algorithmic fairness, with a focus on Computer Science or Law. The search query uses filters to scan through papers based on their title and abstracts. It looks for specific keyword combinations in either of them. The keyword combinations are all variations of terms like “algorithmic fairness”, “fair Machine Learning”, or “discrimination in AI”. By including the wildcard operator (*), we ensured that variations of words are captured that come from the same root (e.g., including the wild card operator before and after ‘fair’ we automatically include terms like “unfair" and “fairly"). Further, we use the (NEAR\5) operator to specify that two words should be placed within a distance of 5 words in the text. The search query was based on an iterative process, adding or removing terms depending on how many search results we obtained. For instance, initially, the query accounted for terms like “bias in Machine Learning”. However, as “bias” is also a purely mathematical (and not ethical) related concept, this yielded too many results, and we excluded this term. After finalising our search query, we conducted a sanity check to ensure that it captured highly cited and well-known papers. We used variations of the same query for the database of papers from ACM FAccT and AAAI AIES proceedings, as well as Hein Online. We chose the first two, as they are the the most prominent conferences on ethics in socio-technical systems. We chose the latter because it is a database containing mostly legal sources, underrepresented in the results of Web of Science.

5.3.2 Selection of Papers

Once we executed our initial search query, we received a total of 6027 papers that required screening for their relevance to the topic of algorithmic fairness. To perform the screening, we utilised Rayyan.ai and established various inclusion and exclusion criteria. To be included in the review, sources were required to have an abstract to ensure that each source under consideration had a minimum level of information available. Moreover, several categories of sources were excluded from the outset. These included introductory notes, book reviews and tutorials, as they were not expected to provide in-depth research content and were not aligned with the intended study scope. Additionally, abstracts of workshops and tutorials for which the full article or chapter could not be accessed were



```
((((TS=(*fairness NEAR/5 algorithm*OR *fairness NEAR/5 machine learning OR
        *fairness NEAR/5 ML OR *fairness NEAR/5 artificial intelligence OR
        *fairness NEAR/5 AI OR  *fairness NEAR/5 classif*
        OR
        *fair NEAR/5 algorithm* OR *fair NEAR/5 machine learning OR *fair NEAR/5 ML OR
        *fair NEAR/5 artificial intelligence OR *fair NEAR/5 AI OR
        *fair NEAR/5 classif*
        OR
        discrimination NEAR/5 algorithm* OR discrimination NEAR/5 machine learning OR
        discrimination NEAR/5 ML OR discrimination NEAR/5 artificial intelligence OR
        discrimination NEAR/5 AI OR discrimination NEAR/5 classif*
        OR
        *justic* NEAR/5 algorithm* OR *justic* NEAR/5 machine learning OR
        *justic* NEAR/5 ML OR  *justic* NEAR/5 artificial intelligence OR
        *justic* NEAR/5 AI OR *justic* NEAR/5 classif*))

         AND SU=(Law OR Computer Science)) NOT SU = (Biology)))
```

Figure 5.1: The Web of Science search query to capture relevant literature, based on key phrases in papers' title and abstract

excluded. Lastly, language was an exclusion criterion, with sources not in English being excluded. We then used the articles' titles as primary indicators of their relevance to the field of algorithmic fairness. In case of ambiguity, we also used the papers' abstracts to decide on their relevance. Through this selection process we ended up with a total of 1570 sources to be included in this scoping review.

5.3.3 Data Extraction

For each of the papers we included for this analysis several features were available, namely their title, abstract and year of publication. Many papers also had a DOI available, which we used to automatically extract additional information from them using `pybliometrics` [112]. This python library utilizes an API to extract information from the Scopus database. In our case, we extracted the names of the papers' authors, and for each author their affiliation at the time of writing the paper (consisting of the name of their institution as well as the corresponding country). This information would be used for answering *RQ4*. To answer parts of research questions 1-3, we also extracted the main expertise areas of each author, as they had self-reported in Scopus.

To extract information on authors' affiliation and expertise on papers without a DOI, we carried out a manual labelling process. We manually checked the papers to extract the authors' names and their affiliations at the time of writing. To determine their area of expertise, we used platforms such as Google Scholar, LinkedIn, and Research Gate. It is important to note that the different labelling processes of authors' expertise may have introduced some errors or biases in our final dataset of papers. This is because the authors' self-reported areas of expertise may differ from the ones we could establish ourselves through a basic web search. Therefore, any results that pertain to this aspect should be regarded as a proxy. Further, many papers had authors coming from mixed backgrounds, with at least one author listing both "Computer Science" and "Law" as their main expertise. Though generally, it could be interesting to inspect contributions from authors with such mixed backgrounds, our result analysis focuses on the work coming only from Law or only Computer Science expertise. This choice was made, because we found many of the "mixed" expertise labels to not be completely reliable, i.e. we found that a lot of Computer Scientists listed "Law" as one of their backgrounds, mostly because "algorithmic fairness" is a topic with some legal implications, not because their work specifically deals with any specific legislation or other legal considerations.

To analyse papers' domain-, demographic and technological focus, as we address in RQ1-RQ3, we manually annotated papers according to these characteristics, using their titles and abstract. We acknowledge that reading full papers would yield more precise results, but since our database consisted of more than 1500 papers, time constraints did not allow this. Through an iterative process, we identified recurring themes regarding the three dimensions and merged similar categories into broader ones, where needed. For example, to describe the papers' technological focus we first had a separate category for "Face Recognition", but because not many papers focussed on this topic, we decided to include them in the broader category "Computer Vision".

Below we list the annotation labels we ended up using for each of the three papers' features:

- **Domain** - Criminal Justice, Education, Employment, Finance, Health, Judicature, Public Sector, Other, None
- **Demographic Groups as Based on** - Age, Disability, Gender, Intersectional, Race, Other, None
- **Technological Approach** - Computer Vision, Data Collection, Hybrid Human-AI, NLP, Resource Allocation, Social Networks, Unsupervised Learning, Ranking, Recommendation, Classification, Other, General

In the result analysis it will become clear that a lot of our found papers do not focus on a specific domain or demographic group (as denoted by the "None" label for either of both features). It is important to note, that both features were only assigned a "non-None" label if papers made some demographic group or some domain the specific focus of their research. To exemplify, many papers introduce novel bias mitigation methods for classification tasks and test their method among others on the COMPAS dataset. Even though this dataset falls under the criminal justice domain, these papers were not tagged as such, unless they specified in their abstract that they went beyond the general benchmark evaluation on this dataset, e.g. consulting domain experts' opinions on the matter or considering domain-specific legislation. Similarly, many papers consider "sex" or "race" as sensitive attributes in their experimental settings. Again, their demographic focus was not tagged as such, unless they dived into specific, historically- or culturally grounded discrimination of those groups. Regarding the technological approach of papers, the "General" label was used if a paper provided a literature review on algorithmic fairness or discussed this as a broad phenomenon, considering many different algorithmic approaches. Also, if a paper's technological approach fell into multiple categories, we chose the more specific one as the primary focus. For instance, a paper on hate speech classification was labelled as "Natural Language Processing" instead of "Classification".

Lastly, to provide labels for the geographical content of papers to answer RQ4, we checked if they mentioned any specific region ("Europe") or country (e.g. "United States") in their title or abstracts and annotated them accordingly.

Common papers in “Other" category: Policing (12), Social Media (8), Sharing Economy (7), Advertisement (6), Insurance (3)

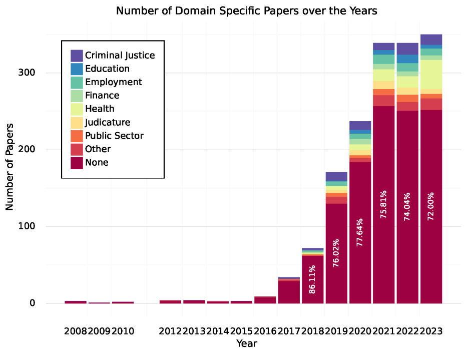


Figure 5.2: The domain focus of papers over the years

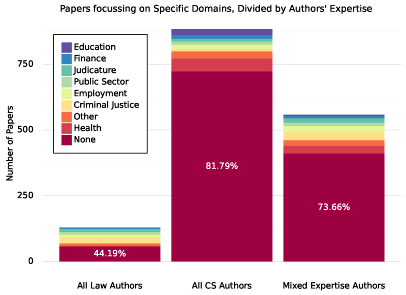


Figure 5.3: The domain focus of papers split by first author’s expertise

Figure 5.4: Over the years papers have become more domain-specific and authors from a legal- or mixed background are more likely to write domain-specific papers. The most discussed domains are health, criminal risk assessment and employment

5.4 Results

After selecting and annotating our papers we conducted the analyses to answer our research questions as outlined in the Introduction of this paper. In the following sections, we will describe the results of these analyses. For each research question, we will first provide a high-level overview of the results, highlighting the research trends related to the specific areas. After, we highlight some specific case studies belonging to less popular research areas, emphasizing the need to dedicate more research to them.

5.4.1 RQ1: The Need for Domain-Specific Approaches

Looking at Figure 5.2, we observe a rising trend in the number of domain-specific papers over the years. Whereas in 2016 only 1̃2% of papers looked at algorithmic fairness through a domain-specific lens, this has risen to 28% by 2023. The most prominent domains revolve around health, criminal justice, judicature and employment. Perhaps unsurprisingly, the rising interest in these domains coincides with case studies within those spheres that have gained public attention. For instance, in 2016 ProPublica published their article on the infamous COMPAS case, an algorithmic risk assessment tool that displayed racial biases against African Americans [83]. In the years following the publication, there is a notable rise in papers addressing fairness in criminal risk assessment. Similarly, we observe an increased interest in fairness in the employment sector after information was released in 2018 about a recruitment tool that Amazon scrapped because of its sexist preference towards male candidates [39]. As we will argue in the next paragraphs, domain-specific approaches open up doors to not just view algorithmic fairness as a general problem, but approach the topic with awareness of the unique challenges in each domain. This holds both when focussing on algorithmic fairness from a legal and technological point of view.

When inspecting Figure 5.3 it is striking how legal authors are much more likely to take this approach to algorithmic fairness than Computer Scientists. A common theme that is touched upon by them, is the adequacy of existing laws to address changes brought by the ubiquitous use of AI systems in different domains. For example, Hertza examines the regulatory landscape in the United States on credit lending, focusing on the Fair Credit Reporting Act (FCRA) and the Equal Credit Opportunities Act (ECOA) [63]. He argues that these laws are inadequate in safeguarding the rights of credit consumers in light of the increasing reliance on big data and advanced algorithmic systems for lending decisions. For example, the FCRA gives consumers the right to access their credit report records, consisting of information about their loan history, on which credit decisions were traditionally based. However, given that the FCRA was enacted in the 1970s, it did not account for the type of third-party data that banks increasingly use to make their decisions, such as lenders' social media profiles or web browsing history. Lacking the right to access this information and understanding how algorithms utilize it, makes it impossible for consumers to challenge algorithmic decisions and assess their fairness. To make up for these gaps in the legislation, Hertza proposes the adoption of the EU General Data Protection Regulation (GDPR) for reforming consumer credit regulation in the US. Because the GDPR is an industry-agnostic framework, it gives individuals the right to access any personal information being processed about them, not just the information on their credit history, that comes from financial institutions. Studies like these highlight the advantage of domain-specific approaches when discussing algorithmic fairness from a legal perspective: as many domains come with their own set of laws, only specific contributions can give insights into their adequacy and invite researchers to challenge them.

Figure 5.3 shows us that Computer Scientists are less likely to take domain-specific approaches. Still, when investigating some of their works, it becomes apparent why specific contibutions are needed to understand the technological challenges within different sectors. Take for instance Pena et al.'s paper set in the employment domain, exploring the type of algorithms typically used to analyse resumes or other professional profiles (e.g., LinkedIn data) in a hiring context. Different from the typical data in other domains, resumes are usually multimodal, as they consist of structured data (e.g., standardized formats to display a person's educational history), unstructured text (e.g., personal biographies) and even images (e.g. profile pictures). Consequentially, automatized solutions for hiring decisions are also multimodal, meaning that one or more multiple models are built to analyze the different data types and base decisions on them. While biases in text-processing or computer vision models have been studied in isolation, the combination of these models and how this combination contributes to new discriminatory biases is less well studied. Another employment-specific study by Rhea et al. even showed how in a hiring setting, simple changes, like whether a resume is processed as raw text or in a PDF file, can change the output of such decision-making systems [108]. We argue that domain-specific approaches are much more likely to reveal problems like these, as they encourage researchers to consider the input data and algorithmic systems that are already in use, rather than making generic assumptions about them.

A final argument for more domain-specific approaches is that they can foster collaborations with interdisciplinary researchers, industry and public institutions, allowing for more in-depth and realistic analyses of unfair practices. Take for instance the study by Elzayhn et al. [44], which is a collaboration between computer scientists, economic- and legal researchers, as well as employees of the US Office of Tax Analysis. They analyse a real-life dataset of taxpayers in the US and - taking domain knowledge about the Tax Payment System into account - analyze if tax audit rates (that partly depend on algorithmic decisions) differ for black and non-black taxpayers. In working with realistic data, the researchers have to deal with challenges often ignored in algorithmic fairness literature, e.g. how to conduct an audit when information about sensitive attributes is not available, but needs to be accurately inferred from the data. Further, by collaborating with the Tax Analysis Office they identify possibilities in reducing the found racial impact, while accounting for their budget and time constraints. Again, this paper forms a contrast to more generic work on algorithmic discrimination, where computer scientists often work in isolation of the institutions using algorithmic systems [126]. In those papers, researchers also commonly use benchmarking datasets for testing their algorithms, which are publicly available datasets, meant to standardize how algorithmic performance is assessed. Though these datasets have their merits, researchers have warned about their quality and the extent to which they can mirror realistic industry use cases [40]. Additionally, only using benchmark datasets can increase the risk of overgeneralizing results obtained from them. Hence, working with domain-specific data that comes from interdisciplinary/industrial collaborations can provide more realistic views on the suitability of AI technologies aimed at addressing algorithmic biases.

To conclude this section, we believe that generic work has been and can still be useful to lay the foundation for algorithmic fairness, but that having more domain-specific case studies will help in tackling more realistic challenges. We have observed a clear trend towards researchers publishing more domain-specific papers, however, as Figure 5.2 shows, many papers remain generic and others revolve around similar domains like health and criminal justice. This may come at the risk of ignoring the risk of algorithmic discrimination in other domains, such as policing, insurance or sharing economy platforms. Hence, broadening the scope of the research field and keeping up to date with the diverse industries/institutions using algorithmic systems, will be essential to reveal the technological challenges and legislation gaps specific to each domain and create tailor-made solutions for them.

5.4.2 RQ2: Making Diverse (Intersectional) Demographic Groups the Focus of the Research

Over the past few years, there has been a slight trend towards publishing more papers that focus on specific demographic groups (e.g., based on race or sex) rather than tackling algorithmic fairness from a generic perspective (see Figure 5.5. In 2023, around 10% of the papers made some demographic group a focus of their work, compared to only 2-6% in 2017-2019. The most prominent categories that papers focus on are race and gender. Also, noticeably, over the years more papers focused on fairness for intersectional groups. Whereas the first two papers on intersectional discrimination appeared only in 2019, in 2022 and 2023, a combined number of 18 papers have focused on this topic.

Similarly, when it comes to focusing on domains when studying algorithmic fairness, we

Common papers in "Other" category: Economic (6), Nationality (5), Political Opinion (2), Sexuality (2), Religion (1)

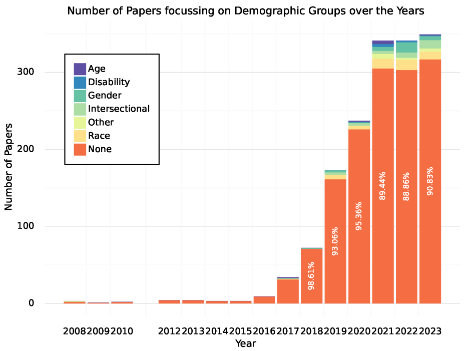


Figure 5.5: The demographic focus of papers over the years

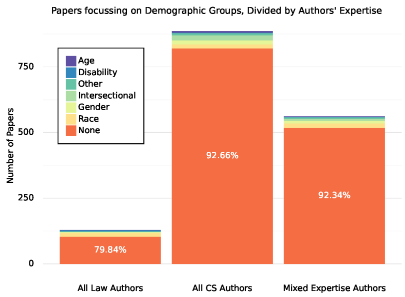


Figure 5.6: The demographic focus of papers split by first author's expertise

Figure 5.7: Most literature tackles algorithmic fairness from a generic perspective, not taking the harms faced by different demographic groups into account.

believe that focusing on demographic groups comes with the advantage of accounting for the historical context and the social dynamics behind discriminatory practices. This viewpoint echoes the argument presented by Hu & Kohler-Hausmann in their paper on algorithmic gender discrimination (2020). Using the example of a decision-making system for college admissions, they highlight how viewing gender as one physical feature in isolation from other attributes, does not do justice to the broader societal implications that come with it. For instance, in college admissions, there are differences across genders in what college programs they apply for or how people's gender shape their opportunities in life. Hence, any fairness considerations made in such a setting should not just consider questions like "are admission ratios equally distributed across sexes?", but also consider why women are less likely to apply for science departments or how societal expectations shaped their past educational and extracurricular activities. Only when recognizing what (algorithmic) fairness would mean in the light of these demographic-specific considerations, meaningful steps can be taken to tackle historic inequalities.

While it is encouraging, that some more papers have taken these demographic-specific approaches over the years, the relatively strong focus on gender and race comes with the risk of ignoring other groups that are targets for discrimination. For instance, over 15 years we only found 8 papers focusing on algorithmic discrimination faced by people with disabilities. While one might argue that general research on algorithmic discrimination is also applicable to ableism, some papers argue why more specialised research is necessary [14, 20]. One argument is that the range of different disabilities is much broader than the range of other sensitive attributes. People with different types of disabilities can be affected by algorithmic systems in vastly different ways. To illustrate, disabilities can range from physical impairments (e.g. being in a wheelchair) to medical conditions (e.g. having cancer), to vision, hearing or cognitive impairments and to psychological conditions (e.g. having depression) [14]. From a technological point of view, Buyl et al. use the example of job recruitment to point out how these distinct categories of disabilities affect algorithms differently (2022): a person with visual impairments, for instance, may need more time on an automated recruitment test, lowering their chances of making it to the next round of a selection procedure. A person with a history of psychological/medical conditions may not have this problem but may instead have bigger gaps on their CV that may be penalised by an algorithm. Lastly, automated video analysis software, used, e.g. for job interviews, may perform okay on either of both groups but not on people with speech impairments. The same paper then discusses the idea of “reasonable accommodation" as a possible technical solution to address these problems: if algorithmic systems have information on the type of disability of any individual, they can be designed to accommodate each of them. For instance, automated video analysis software could be designed to process sign language to accommodate people with speech impairments. Additionally, algorithms for analysing CVs could be designed to not penalize career gaps if a job applicant has a history of medical conditions. While these are reasonable adjustments from a technological perspective, contributions from a legal perspective point out how difficult it may be to gather data about peoples' disabilities, as people might prefer not to disclose this information, for fear of it being abused or lack of discretion in handling the data [14, 20]. A paper by Binns & Kirkham (2021), therefore, explores the role of data protection and equality law, in ensuring algorithmic fairness for disabled people, while simultaneously protecting their privacy. For instance, they highlight how data protection laws (e.g., the GDPR) allow institutions to collect “special category data" (including information about persons' disability status) if they have an appropriate lawful basis for wanting to process this data. Further, they emphasize how these laws can create a safer and more trustworthy environment around sharing personal data, as they define clear boundaries regarding how the data should be used and with which parties it can be shared. Hence, ensuring strict enforcement of these laws can increase peoples' willingness to share sensitive information and ensure that this information is only used to provide “reasonable accommodation", as mentioned earlier.

The example papers surrounding algorithmic fairness for disabled people illustrate the importance of delving into specific demographic groups to gain a clearer understanding of how they are affected by algorithmic systems. By outlining both technologically- and legally-driven research papers, we emphasize how expertise from both disciplines is needed to find realistic solutions for the addressed challenges. Inspecting Figure 6, this is especially a call for Computer Science researchers to adopt such demographic-specific approaches, as they are less likely to do so than their legal counterparts. Specifically, only around 7% of Computer Scientists make specific demographic groups the main focus of their research, while this ratio is 14% higher for Law experts. Further, we emphasize again, how the research on algorithmic fairness needs to broaden its scope and include various demographic groups that go beyond just the race and gender of people. As Figure 5.7 points out, there are still many demographic features that are barely considered in current research efforts, posing the risk of overlooking the harms faced by diverse and intersectional communities.

Common papers in "Other" category: Generative AI (13), Federated Learning (11), Speech Recognition (10), Price Discrimination (8), Representation Learning (8), Regression (7), Internet of Things (6)

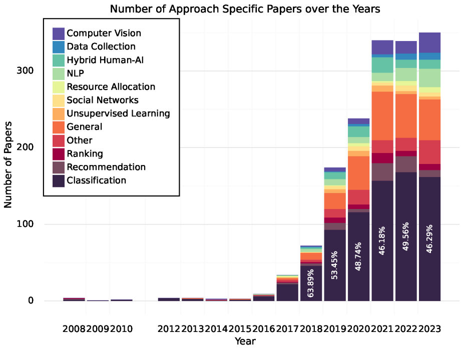


Figure 5.8: The technological focus of papers over the years

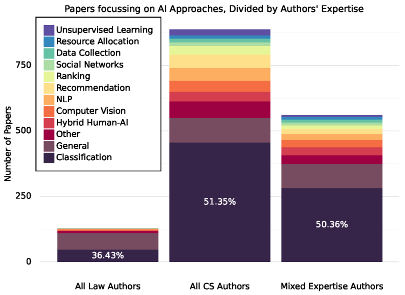


Figure 5.9: The technological focus of papers split by first author's expertise

Figure 5.10: Classification remains the prime technological focus of studies on algorithmic fairness. Computer Science research is slowly getting more diverse in their addressed technologies.

5.4.3 RQ3: Moving beyond Classification

In Figure 5.8, we display how focuses on different approaches have developed over the years. What is striking, yet unsurprising, is that over the years "Classification" remains the most discussed technology regarding algorithmic fairness. While on the one hand, this may be a result of our generic search query, which did not include specific terms like "Clustering" or "Speech Recognition", it undoubtedly reflects an already known concern, that researchers often do not look beyond fairness in classification tasks based on tabular data [28, 65, 109]. Still, we observe a trend that the range of discussed technologies gets wider and more diverse over the years, as in 2018 still about 64% of all contributions focussed on classification while by 2023 this has gone down to 46%. The most heavenly discussed technologies besides classification are computer vision, NLP and recommendation systems.

When inspecting where more diverse discussions on AI technologies come from in Figure 5.9, we immediately see that authors from a pure Law background are the least diverse, with nearly all their contributions discussing classification tasks or "AI" as a general phenomenon. Partly, this may result from the generic ways in which AI legislation is phrased. Since technology develops so rapidly and unpredictably, it is impossible to account for all its potential forms. Hence generic guidelines around its usage allow policy writers to encompass more use cases, likely reflected in the scientific literature about these guidelines. Still, neglecting the precise shapes that algorithms can take can lead to an incomplete understanding of their usage and substantial gaps in the laws regulating them. This is exemplified in one of the legal contributions from Keunen (2023). She investigates data collection practices around tax audits and the extent to which they can be considered “fishing expeditions". Specifically, she examines their privacy-intruding nature, wherein an excessive amount of taxpayer information is collected and analyzed in the pursuit of detecting fraud, before having sufficient justification for why these taxpayers are targeted as potentially fraudulent and why extra data needs to be collected for them. In her work, Keunen alludes to various technologies used to collect this data, namely automated web scraping and web crawling algorithms. While she primarily raises privacy concerns related to these practices, it is clear how also from an algorithmic fairness standpoint these techniques can be highly problematic. For instance, another work by Jo & Gebru, explains how the availability and nature of data that can be crawled from online spaces is influenced by demographic factors (2020), with e.g. younger generations being more represented on the internet than older ones. Consequently, fraud-detection algorithms relying on web-crawled data may disproportionately impact younger groups, as more potentially incriminating data is available about them. Despite the clear fairness and privacy concerns around web crawling, Keunen (2023) points out that their regulation and the extent to which they can be considered “fishing expeditions" is still unclear: explicit legislation is not available and so far only case law serves as an indication for which data collection practices are prohibited. Hence, Keunen’s work showcases, how for identifying other gaps in the legislation related to privacy and algorithmic fairness, more legal experts need to dive into specific technologies, rather than primarily focusing on AI as a general problem. Collaborating with Computer Science experts in doing so, will be important to stay on top of the fast-paced development of technology.

To highlight some of the complex fairness considerations, that Computer Scientists currently make about other non-classification technologies consider the work of Jalal et al. (2021), who explore image-reconstruction algorithms. These algorithms take low-resolution images as input and try reconstructing them into higher resolutions. In doing so, they are known to be biased. For instance, when low-resolution images of a black person are given as input they are likely to reconstruct it into the image of a white person. The work addresses the intricacies of even defining “fairness" in such a setting. Unlike classification tasks, where a classifier’s decision should be independent of sensitive attributes (e.g., employment decisions should not be influenced by race), fair image reconstruction algorithms must produce outputs that align with the sensitive characteristics in the input. This introduces the challenge of estimating race and other sensitive attributes from images, a task complicated by their non-discrete and highly ambiguous nature. Another example of non-trivial fairness issues concerns the use of Generative AI systems. While our scoping review found only 13 papers related to this technology, it seems reasonable to assume that this number will rise, given the popularity of ChatGPT, DallE, and other generative systems. Venkit et al. (2023) are some of the few authors exemplifying the fairness issues arising through these systems, examining how text generation models exhibit different sentiments and toxicity levels depending on the nationalities they are prompted to write about. For example, when prompted to write about Irish people, human annotators perceived the articles to be mostly benign and generic, while texts about Tunisian people were rated to be much less positive and more focused on negative events in the country. How such texts can perpetuate harmful stereotypes and how to restrict these models are still largely unexplored questions. The topic is complicated considerably by the seemingly infinite topics these systems can be prompted to write about and all the ways a chatbot-human interaction could unfold. Hence, for just defining what it means for such a huge system to be fair, more technological and legal research is necessary.

While it lies outside the scope of this paper to discuss the fairness concerns arising in

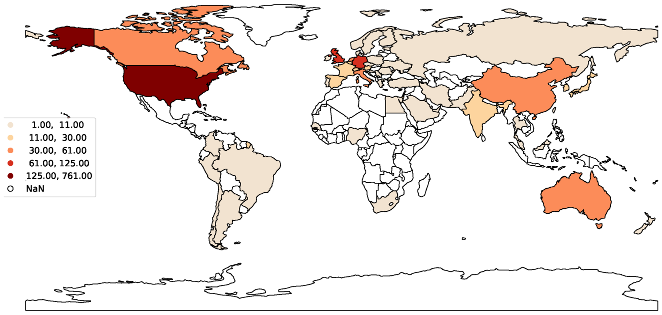


Figure 5.11: A world map displaying the affiliation countries of each paper's first author

all other kinds of algorithmic systems, it should be clear that they can go far beyond the matters addressed in the typical classification setting. As technology advances rapidly and various algorithmic systems become more prevalent among the public, it is clear that researchers should make an effort to keep up with this development and extend their focus beyond the conventional realms.

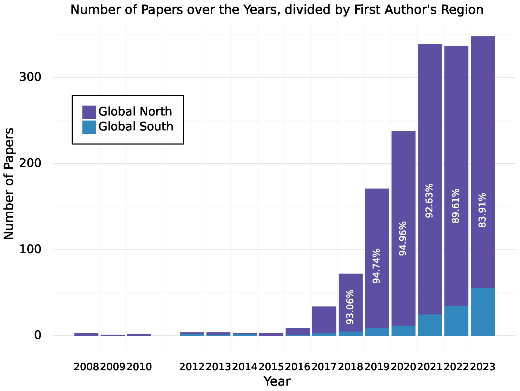
5.4.4 RQ4: Considering Global Perspectives

To analyse the geographical context in which algorithmic fairness was discussed, we both examined the authors' affiliation countries as well as the geographic focus in papers' content. For the former analysis, we considered the first authors as the primary contributors.

5.4.4.1 Authors' Affiliation

In Figure 5.11, we display a geographical heatmap, displaying the number of papers divided by each paper's first author's affiliation country. At first sight, it is evident that most contributions come from authors affiliated with institutes in North America and Europe, and papers from authors affiliated with other countries are quite sparse. To further investigate this trend, we classified each first author's affiliation country according to whether it belongs to a Global North or Global South country[3]. In Figure 5.1a, we display how this geographical context of the first authors has developed over the years. From this Figure, we see that consistently most contributions come from authors hailing from the Global North. While this predominant presence persists, a noteworthy shift is discernible over the years, with an almost 20% uptick in contributions from the Global South institutions in 2023, signalling a gradual re-balancing compared to earlier years where only about 5% of contributions come from the Global South. As we will see in the following section, this shift is crucial for challenging the Northern-centric perspective within AI research.

[3] We used UNCTAD's classification in which Global North is understood as countries in Europe and Nothern America; and including Israel, Japan, Australia and New Zealand. The Global South consists of countries in Africa, Asia, South America and the Caribbean. `https://unctadstat.unctad.org/EN/Classifications.html`



(a) Though the majority of papers come from authors affiliated in the global north, contributions from papers from global southern affiliations are rising

Country/Regional Focus	# Papers
United States	82
Europe	60
United Kingdom	12
China	9
India	7
Australia	6
Canada	4
Brazil, Netherlands, Germany, Italy, Africa	3
Spain, Singapore, France, Austria, Russia, Chile, Global South	2
New Zealand, Switzerland, Uruguay, Mexico, South America, Bangladesh, South Korea, Maldives, Vietnam, Philippines, Japan, Asia, Israel, United Arab Emirates, Nigeria	1

(b) Number of times different countries/geographical regions were made the focus of a papers' content

5.4.4.2 Geographical Focus in Papers' Content

Next to the first authors' affiliation country, we analysed the papers' geographic focus, as estimated by them mentioning any specific country/region name in their title or their abstract. In Table 5.1b we display the results of this analysis.

Intriguingly, the results unveil a similar representation gap, as we have found upon examining the authors' affiliation, with most papers concentrating on countries in the Global North. There could be several reasons for the under-representation of work from/about the Global South, especially those from Africa:

- Databases may not systematically include publications from the Global South, hinting at access challenges or a predilection for regional databases.
- The search query methodology, requiring specification of the Global South country in the title, may inadvertently limit the breadth of results.

These, among other reasons, such as the limited resources in research institutions in the Global South, language barriers, and lack of engagement with literature from the South, contribute to the underrepresentation of certain geographical regions [97].

Nevertheless, as already mentioned, the past years have observed a rise in both papers coming from non-northern institutions, as well as a rise of papers concentrating on algorithmic fairness in global southern countries. Notable examples include studies on Predictive Policing in New Delhi [92], Early Prediction of At-Risk Students in Uruguay [107], and discussions on Algorithmic Fairness in China and Brazil [103, 132]. These papers delve into the intricacies of AI applications within diverse cultural, economic, and legal contexts, emphasising the need for nuanced considerations in algorithmic development. The paper "The Algorithmic Imprint " by Eshan et al. (2022) provides an especially clear example of why these kinds of considerations are necessary, and why it is important to have more diverse and inclusive voices in the narrative about algorithmic fairness. The paper discusses the grading algorithm developed to predict students' A Level results when the exams could not be administered due to the Covid19 pandemic. Though there were many unfairness complaints about the algorithm's predictions (and they were ultimately discarded) they turned out to be especially unfair towards students from commonwealth schools outside of the UK, in which A-Levels are also the primary form of examination. By focussing on the specific case of Bangladeshi schools, the authors find how UK-based assumptions throughout the algorithm's development, can explain the disparity in predictions. One example is, that the grade predictions were based on performance in mock exams, assuming that good performance on a mock exam is predictive of a good performance on the real one. While intuitively this might make sense, this assumption neglects the learning culture in Bangladesh where much more emphasis is put on final examination and mock exams are not a common part of the curriculum. To have some data to work with, Bangladeshi students were forced to take some hurriedly set up tests, which they were not used to and had little time to prepare for. Needless to say, grade predictions based on this type of data, did not reflect students' real capabilities. Another flaw in the design of the algorithm was the decision to base grade predictions on the historical performance of the student's school. If such data were not available, international averages were used instead. This proved especially problematic for Bangladeshi schools, which were less likely to possess (digital) records of historical performance. Consequently, predictions frequently leaned on the international averages, even though these were lower than the (unrecorded) actual historical performances. These are only a few of many examples, of how a lack of cultural considerations led to an algorithm, that was ultimately more unfair to some geographic groups than others.

Having additional papers adopting a cultural- and geographic-specific approach can contribute to a more diverse and comprehensive understanding of algorithmic fairness, shedding light on various perspectives and mitigating unfairness across different regions. In addition to specific case studies, we also found several papers contributing to a more global discourse on algorithmic discrimination. For instance, Amugongo et al. (2020) examine fairness from a philosophical standpoint, exploring how African-based "Ubuntu ethics" can enrich discussions about the essence of fairness. Another example is Nwafor's paper (2021), which delves into the policies and laws from global southern countries concerning AI systems. Studies like these will be essential to make sure that legislation outside Europe and the US is ready for upcoming technological developments. In her paper Nwafor also advocates for a diverse representation in AI's design, development, deployment, and governance. Neglecting to engage marginalised communities in AI's development, leads to technological innovation being based on a a narrow slice of the world, lacking a comprehensive analysis of diverse global groups. Integrating more diverse perspectives not only enhances our understanding of algorithmic fairness but also emphasizes the importance of cross-cultural learning to create more inclusive and equitable AI systems.

5.5 Discussion & Conclusion

In this paper, we have presented a scoping review of the current literature on algorithmic fairness. We selected and annotated 1570 papers to examine the evolution of the field in terms of their domain-, demographic-, and technological focus and their interdisciplinary nature, while also inspecting the geographical context in which the research takes place.

We acknowledge two major limitations in our analysis. First, we used a basic search query to collect papers for our dataset by using general terms such as "machine learning" and "AI". However, this approach did not account for more specialised terms regarding papers technological approach, their domain or demographic focus, which may have caused us to miss valuable contributions within those areas.

The second limitation concerns our manual annotation process, which we based on the papers' titles and abstracts, rather than their full text. While both should give a good reflection on the paper's main topic, some important nuances might have been missed.

Despite these limitations, our results still provide a valuable overview of the current research landscape surrounding algorithmic fairness, in particular the trending topics and the gaps within the field. Our analysis shows that over the years, research has started focusing on more specific and a wider variety of topics in terms of the addressed technologies, domains and demographic groups. This stands in contrast to early work in the field, which discussed algorithmic fairness concerns solely in classification tasks, without questioning domain-specific challenges or the harms different demographic groups might face. Through highlighting some papers, we have made a case for why more specialised research is necessary, both from a legal and technological point of view, as non-specific approaches come at the risk of ignoring the algorithmic systems that are actually used by companies and the unique considerations that go into tackling their discriminatory behaviour.

Finally, we examined the geographical context of ongoing research, by analysing authors' affiliations and the papers' geographical focus. While the trend is slowly changing, most papers come from global north countries and focus on the algorithmic development and regulations there. Through some case studies, we have emphasized how a lack of diverse cultural considerations in developing algorithms, can lead to severe discriminatory results depending on where they are applied. Therefore, an inclusive approach is necessary to comprehend the broader implications of algorithmic fairness in distinct contexts and how to address these.

Chapter 6

Conclusion

This thesis has made various contributions towards shifting the discussion around algorithmic fairness away from single mathematical fairness notions, and instead towards more rigorous and human-centered approaches for measuring and mitigating bias.

Chapter 1 and 2 have focussed on bias detection in ADM systems. Specifically, in Chapter 1 we dealt with the potential of interactive auditing toolkits for conducting rigorous bias audits. We identified the requirements for these toolkits to be usable in realistic settings. We have inspected existing tools and shown how some of them fall short in exposing more complex patterns of bias, such as intersectional discrimination or the potential causes of such biases. By comparing different toolkits we also highlighted how their functionality can be combined, to ultimately design a tool that meets auditors' needs. In Chapter 2 we zoomed in on one specific part of bias audits, namely the measurement of individual fairness. One current individual fairness notion, as defined by situation testing, is based on the principle of *treating likes alike*: if we want to know if an individual of some specific demographic group received fair treatment, we need to compare their decision output to those of other similar individuals of other demographic groups. Again, we have highlighted the importance of adapting context-dependent approaches for measuring fairness, by showing the shortcomings of an arbitrarily defined distance function to measure similarity in this setting. Moreover, we proposed a method to learn an appropriate distance function from the data, such that similarities of individuals are only defined based on attributes relevant to a decision task.

After concentrating on appropriate tools and methods for bias detection in this thesis's first two chapters, we proposed a new bias mitigation technique in Chapter 3. We explored how the framework of selective classification, that allows classification model to not make predictions for uncertain instances, can be extended to account for the unfairness of predictions. In other words, we designed a method for learning a selective classifier that rejects predictions that are either uncertain or unfair. We ensured that the rejection-mechanism for unfair predictions is completely explainable and highlighted how these explanations, that show why an original prediction was seen as discriminatory, can empower humans to make more well-informed decisions on these instances. With this contribution we highlighted the importance of properly understanding an ADM model's biases before these biases are resolved.

In Chapter 4 we moved from algorithms for bias mitigation to evaluating their effectiveness. We made a case for why the popular evaluation scheme, provided by the fairness-accuracy trade-off is suboptimal for two reasons: first, this scheme measures fairness only through single mathematical definitions, rather than encouraging more thorough bias audits. Second, this evaluation scheme is logically flawed, as it requires a high accuracy on decision labels that we believe to be flawed and biased to begin with. As an alternative way to benchmark fairness interventions, we have presented a new dataset containing a biased and fair version of its decision labels, providing full information about which instances are discriminated. We have shown how this dataset can be used to evaluate fairness interventions, by applying them on the biased version of the labels and testing their accuracy on the fair ones. We have also shown how some fairness interventions that work well according to the traditional evaluation scheme, do not necessarily perform well according to this new one, further emphasizing the flaws of the accuracy-fairness trade-off.

In the final chapter of this thesis, we have zoomed out from specific concerns regarding bias detection and mitigation in ADM systems and instead provided a broader overview of the field of algorithmic fairness. Specifically, we presented a scoping review of how the research has evolved in the last 15 years, regarding its geographical context, discussed domains, demographic groups and technologies. Distinguishing between contributions from technological and legal experts, we established popular research trends and identified research gaps that deserve more attention from the community. Based on the findings of this scoping review and based on the content of the other chapters of this thesis we can establish various directions for future research that should evolve from this thesis.

6.1 The Need for Case-Studies

In Chapters 1-3 of this thesis we have advocated for thorough and rigorous methods for detecting and mitigating biases in ADM systems. We need to take a context-dependent approach, that allows us to take the intricacies of a decision task into account and lets us understand which biases are unacceptable and which ones can be justified by the nature of the domain. While we have discussed tools and algorithms that enable such context-driven approaches and empower human auditors to incorporate their domain knowledge into decision tasks, we have not yet tested any of these approaches in a specific setting like hiring or lending. As we have pointed out in our scoping review, conducting case-studies and collaborating with experts from a field is highly important to understand relevant legislation, time- and budget constraints, available data and current technological practices within a domain. Hence, until these kinds of case studies are conducted, it is difficult to estimate under which circumstances our proposed solutions are viable and how they may be further adapted for realistic settings.

For instance, an intriguing case study could involve examining the fair selective classifier introduced in Chapter 3 within a lending context. By testing this classifier in such a scenario, we could observe how bankers review predictions flagged as unfair and how they utilize the explanations provided for rejections. Additionally, this exploration could offer insights into the compatibility of the methodology with the bank's data and workflow and the resources they can allocate to a human-in-the-loop approach. As highlighted in Chapter 5, we argue that it is through case studies like these that we can propel the field of algorithmic fairness forward and discover effective strategies for mitigating the harms caused by real-life ADM systems.

6.2 Involving End-Users

In this thesis, we have taken a technologically-centered approach to algorithmic fairness, exploring how tools and algorithms can enable more rigorous and context-driven bias audits and interventions. However, as mentioned in the introductory chapter, this approach is not meant to undermine the importance of stakeholder-driven approaches in the field.

As the name implies, stakeholder-driven approaches involve the active engagement of various stakeholders in the design, evaluation, and testing of a technological system. Referring back to the example from the previous section, further developing our selective classifier in the context of a lending setting would involve engaging bankers who use the system and its decision subjects. We have already explained how the involvement of bankers is essential for understanding how a system can fit into their current workflow and practices. Equally important is understanding the perspectives of loan applicants, as their lived experiences reveal their unique needs and challenges when applying for a loan.

This also relates to the points we made in our scoping review about focusing on specific demographic groups: understanding both the historical discrimination these groups have faced and how these dynamics may play out under current circumstances is necessary to address any inequalities.

The research presented in this thesis has not engaged in stakeholder-driven design and has not viewed algorithmic fairness from a demographic-specific lens. While we believe our proposed methodology can still serve as a solid foundation for building concrete and contextually tailored approaches, this is certainly an area for future research. Related to this, it is important to look beyond sexism and racism, the most commonly addressed forms of discrimination in this thesis. As mentioned in Chapter 5, there are many more sensitive attributes to consider. Furthermore, as discussed in Chapter 1, algorithmic biases can still affect marginalized groups even if information about their nationality, disability status, or other sensitive characteristics (beyond gender and race) is unavailable.

Exploring how the proposed methodologies in this thesis, such as learning a distance metric for individual fairness (Chapter 2) and building a fair selective classifier (Chapter 3), can be applied to a variety of sensitive characteristics (some of which might not be recorded in the data) would be a worthwhile effort. Adopting a stakeholder-driven approach for this would be highly beneficial in addressing questions such as under what circumstances individuals would feel comfortable sharing sensitive information and understanding how intersectionally defined groups might experience unique harms imposed by an algorithmic system.

6.3 Going from static to dynamic settings

In Chapter 1-4 of this thesis, we have focussed on bias mitigation in automated decision tasks, also referred to as classification tasks. Even if Chapter 5 has highlighted how there are many more algorithmic technologies to consider, it is worthwhile to mention how even the heavily discussed classification setting tends to be more complex than discussed in this thesis. This is largely due to the dynamic nature under which those systems operate. So far we have only considered static settings, assuming that after the data for some task is gathered, a classification model is trained on a part of this data and its biases are audited for on another part. Once a classification model (with or without fairness interventions) is selected, we assume it does not change throughout its deployment and is used in isolation from other algorithmic systems. In reality, both the learning and deployment process of a decision-making model are a lot more complex: throughout its deployment, a model may be retrained on new batches of data several times. What is more, the nature of the data used for retraining might be directly influenced by the decisions the system has made in the past [126]. For instance, in a lending setting a bank continuously gathers new information about which loan receivers manage to pay back their credits, which can be used to fine-tune decisions for new applicants. However, since information is only available on those individuals who were granted a loan by the ADM system in the first place, new biases might be introduced as a result of this retraining process. Additionally, decision-making models might be used in combination with other algorithmic systems, and the incoming data of new applicants may change over time as a result of concept drift [28].

It is clear that fairness is even less straightforward to address in highly dynamic settings like these. Hence, it remains to be studied how the tools and algorithms discussed in this thesis can be adapted to account for this non-static nature. Similarly as pointed out in the previous sections, we believe that case studies, working with realistic data and investigating the current practices within a domain, are key to shed light on these concerns.

Bibliography

[1] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *VLDB*, pages 487–499. Morgan Kaufmann, 1994.

[3] Y. Ahn and Y.-R. Lin. Fairsight: Visual analytics for fairness in decision making. *IEEE transactions on visualization and computer graphics*, 26(1):1086–1095, 2019.

[4] A. Artelt, J. Brinkrolf, R. Visser, and B. Hammer. Explaining reject options of learning vector quantization classifiers. In *IJCCI*, pages 249–261. SCITEPRESS, 2022.

[5] A. Artelt and B. Hammer. "even if ..." - diverse semifactual explanations of reject. In *SSCI*, pages 854–859. IEEE, 2022.

[6] A. Artelt, R. Visser, and B. Hammer. "i do not know! but why?" - local model-agnostic example-based explanations of reject. *Neurocomputing*, 558:126722, 2023.

[7] E. Ash, N. Goel, N. Li, C. Marangon, and P. Sun. WCLD: curated large dataset of criminal cases from wisconsin circuit courts. 2023.

[8] Automated Employment Decision Tools. Automated employment decision tools. `https://rules.cityofnewyork.us/rule/automated-employment-decision-tools-updated/`, 2023.

[9] M. Bao, A. Zhou, S. Zottola, B. Brubach, S. Desmarais, A. Horowitz, K. Lum, and S. Venkatasubramanian. It's compaslicated: The messy relationship between rai datasets and algorithmic fairness benchmarks. *NeurIPS 2021 Track on Datasets and Benchmarks*, 35, 2021.

[10] S. Barocas, A. Guo, E. Kamar, J. Krones, M. R. Morris, J. W. Vaughan, W. D. Wadsworth, and H. Wallach. Designing disaggregated evaluations of ai systems: Choices, considerations, and tradeoffs. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 368–378, 2021.

[11] S. Barocas, M. Hardt, and A. Narayanan. Fairness in machine learning. *Nips tutorial*, 1:2, 2017.

[12] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

[13] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

[14] R. Binns and R. Kirkham. How could equality and data protection law shape ai fairness for people with disabilities? *ACM Transactions on Accessible Computing (TACCESS)*, 14(3):1–32, 2021.

[15] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt. 'it's reducing a human being to a percentage' perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems*, pages 1–14, 2018.

[16] A. Birhane, E. Ruane, T. Laurent, M. S. Brown, J. Flowers, A. Ventresque, and C. L. Dancy. The forgotten margins of ai ethics. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 948–958, 2022.

[17] W. Blanzeisky, P. Cunningham, and K. Kennedy. Introducing a family of synthetic datasets for research on bias in machine learning. *arXiv preprint arXiv:2107.08928*, 2021.

[18] K. Boyd. Designing up with value-sensitive design: Building a field guide for ethical ml development. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2069–2082, 2022.

[19] C. S. Brown and E. A. Stone. Gender stereotypes and discrimination: How sexism impacts development. *Advances in child development and behavior*, 50:105–133, 2016.

[20] M. Buyl, C. Cociancig, C. Frattone, and N. Roekens. Tackling algorithmic disability discrimination in the hiring process: An ethical, legal and technical analysis. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1071–1082, 2022.

[21] M. Buyl and T. De Bie. Inherent limitations of ai fairness. *Communications of the ACM*, 67(2):48–55, 2024.

[22] Á. A. Cabrera, W. Epperson, F. Hohman, M. Kahng, J. Morgenstern, and D. H. Chau. Fairvis: Visual analytics for discovering intersectional bias in machine learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 46–56. IEEE, 2019.

[23] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30, 2017.

[24] G. Casella and R. L. Berger. Statistical inference duxbury press. *Pacific Grove, CA.*, 2002.

[25] L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 319–328, 2019.

[26] A. Chamberlain. Demystifying the gender pay gap: Evidence from glassdoor salary data. `https://www.classlawgroup.com/wp-content/uploads/2016/11/glassdoor-gender-pay-gap-study.pdf`, 2016.

[27] S. Chiappa and W. S. Isaac. A causal bayesian networks viewpoint on fairness. In *IFIP International Summer School on Privacy and Identity Management*, pages 3–20. Springer, 2018.

[28] A. Chouldechova and A. Roth. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, 2020.

[29] C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Trans. Inf. Theory*, 16(1):41–46, 1970.

[30] F. Condessa, J. M. Bioucas-Dias, C. A. Castro, J. A. Ozolek, and J. Kovacevic. Classification with reject option using contextual information. In *ISBI*, pages 1340–1343. IEEE, 2013.

[31] S. Corbett-Davies and S. Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[32] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

[33] C. Cortes, G. DeSalvo, and M. Mohri. Theory and algorithms for learning with rejection in binary classification. *Annals of Mathematics and Artificial Intelligence*, pages 1–39, 2023.

[34] P. Cortez and A. M. G. Silva. Using data mining to predict secondary school student performance. 2008.

[35] S. Costanza-Chock, I. D. Raji, and J. Buolamwini. Who audits the auditors? recommendations from a field scan of the algorithmic auditing ecosystem. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1571–1583, 2022.

[36] K. Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In *University of Chicago Legal Forum: Vol. 1989*. 1989.

[37] L. Darling-Hammond. Unequal opportunity: Race and education. *The Brookings Review*, 16(2):28–32, 1998.

[38] D. Dash and M. J. Druzdzel. Robust independence testing for constraint-based learning of causal structure. In *UAI*, volume 3, pages 167–174, 2003.

[39] J. Dastin. Insight - amazon scraps secret ai recruiting tool that showed bias against women. https://www.reuters.com/article/idUSKCN1MK0AG/, 2018.

[40] F. Ding, M. Hardt, J. Miller, and L. Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

[41] J. Dodge, Q. V. Liao, Y. Zhang, R. K. Bellamy, and C. Dugan. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces*, pages 275–285, 2019.

[42] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[43] R. El-Yaniv and Y. Wiener. On the foundations of noise-free selective classification. *J. Mach. Learn. Res.*, 11:1605–1641, 2010.

[44] H. Elzayn, E. Smith, T. Hertz, A. Ramesh, J. Goldin, D. E. Ho, and R. Fisher. *Measuring and mitigating racial disparities in tax audits*. Stanford Institute for Economic Policy Research (SIEPR), 2023.

[45] L. Enqvist. 'human oversight'in the eu artificial intelligence act: what, when and by whom? *Law, Innovation and Technology*, 15(2):508–535, 2023.

[46] European Commission. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. `https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206`, 2021.

[47] A. Fabris, S. Messina, G. Silvello, and G. A. Susto. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, 2022.

[48] M. Favier, T. Calders, S. Pinxteren, and J. Meyer. How to be fair? a study of label and selection bias. *Machine Learning*, 112(12):5081–5104, 2023.

[49] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[50] L. Fischer, B. Hammer, and H. Wersing. Optimal local rejection for classifiers. *Neurocomputing*, 214:445–457, 2016.

[51] B. Fish, J. Kun, and Á. D. Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM international conference on data mining*, pages 144–152. SIAM, 2016.

[52] W. Fleisher. What's fair about individual fairness? In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 480–490, 2021.

[53] V. Franc, D. Prusa, and V. Voracek. Optimal strategies for reject option classifiers. *Journal of Machine Learning Research*, 24(11):1–49, 2023.

[54] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338, 2019.

[55] A. Gangrade, A. Kag, and V. Saligrama. Selective classification via one-sided prediction. In *AISTATS*, volume 130, pages 2179–2187. PMLR, 2021.

[56] Y. Geifman and R. El-Yaniv. Selective classification for deep neural networks. In *NIPS*, pages 4878–4887, 2017.

[57] Y. Geifman and R. El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *ICML*, volume 97, pages 2151–2159. PMLR, 2019.

[58] N. Goel, M. Yaghini, and B. Faltings. Non-discriminatory machine learning through convex fairness criteria. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 116–116, 2018.

[59] D. Hadwick and S. Lan. Lessons to be learned from the dutch childcare allowance scandal: a comparative review of algorithmic governance by tax administrations in the netherlands, france and germany. *World tax journal.-Amsterdam*, 13(4):609–645, 2021.

[60] K. Hendrickx, L. Perini, D. V. der Plas, W. Meert, and J. Davis. Machine learning with a reject option: A survey. *ArXiv*, abs/2107.11277, 2021.

[61] R. Herbei and M. H. Wegkamp. Classification with reject option. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 709–721, 2006.

[62] R. Herbei and M. H. Wegkamp. Classification with reject option. *Can. J. Stat.*, 34(4):709—-721, 2006.

[63] V. A. Hertza. Fighting unfair classifications in credit reporting: Should the united states adopt gdpr-inspired rights in regulating consumer credit. *NYUL Rev.*, 93:1707, 2018.

[64] A. L. Hoffmann. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, 22(7):900–915, 2019.

[65] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudik, and H. Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–16, 2019.

[66] M. Hort, Z. Chen, J. M. Zhang, M. Harman, and F. Sarro. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing*, 2023.

[67] L. Hu and I. Kohler-Hausmann. What's sex got to do with fair machine learning? *arXiv preprint arXiv:2006.01770*, 2020.

[68] L. Huang, C. Zhang, and H. Zhang. Self-adaptive training: beyond empirical risk minimization. In *NeurIPS*, 2020.

[69] C. Ilvento. Metric learning for individual fairness. *arXiv preprint arXiv:1906.00250*, 2019.

[70] H. Jiang and O. Nachum. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 702–712. PMLR, 2020.

[71] E. Jones, S. Sagawa, P. W. Koh, A. Kumar, and P. Liang. Selective classification can magnify disparities across groups. In *ICLR*, 2021.

[72] N. Kallus and A. Zhou. The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. *Advances in neural information processing systems*, 32, 2019.

[73] F. Kamiran and T. Calders. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*, pages 1–6. IEEE, 2009.

[74] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[75] F. Kamiran, I. Žliobaitė, and T. Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and information systems*, 35(3):613–644, 2013.

[76] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14, 2020.

[77] B. Kavšek and N. Lavrač. Apriori-sd: Adapting association rule learning to subgroup discovery. *Applied Artificial Intelligence*, 20(7):543–583, 2006.

[78] N. Kilbertus, A. Gascón, M. Kusner, M. Veale, K. Gummadi, and A. Weller. Blind justice: Fairness with encrypted sensitive attributes. In *International Conference on Machine Learning*, pages 2630–2639. PMLR, 2018.

[79] M. E. Kite and B. E. Whitley Jr. *Psychology of prejudice and discrimination*. Routledge, 2016.

[80] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[81] J. Kühne, C. März, et al. Securing deep learning models with autoencoder based anomaly detection. In *PHM Society European Conference*, volume 6, pages 13–13, 2021.

[82] R. L. Cardoso, W. Meira Jr, V. Almeida, and M. J. Zaki. A framework for benchmarking discrimination-aware models in machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 437–444, 2019.

[83] J. Larson, S. Mattu, L. Kirchner, and J. Angwin. How we analyzed the compas recidivism algorithm. `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`, 2016.

[84] P.-M. Law, S. Malik, F. Du, and M. Sinha. Designing tools for semi-automated detection of machine learning biases: An interview study. *arXiv preprint arXiv:2003.07680*, 2020.

[85] J. K. Lee, Y. Bu, D. Rajan, P. Sattigeri, R. Panda, S. Das, and G. W. Wornell. Fair selective classification via sufficiency. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 6076–6086. PMLR, 2021.

[86] M. S. A. Lee and J. Singh. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–13, 2021.

[87] D. Lenders and T. Calders. Learning a fair distance function for situation testing. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 631–646. Springer, 2021.

[88] A. Liberati, D. G. Altman, J. Tetzlaff, C. Mulrow, P. C. Gøtzsche, J. P. Ioannidis, M. Clarke, P. J. Devereaux, J. Kleijnen, and D. Moher. The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Annals of internal medicine*, 151(4):W–65, 2009.

[89] B. T. Luong, S. Ruggieri, and F. Turini. k-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 502–510, 2011.

[90] M. A. Madaio, L. Stark, J. Wortman Vaughan, and H. Wallach. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[91] N. Malik and P. V. Singh. Deep learning in computer vision: Methods, interpretation, causation, and fairness. In *Operations Research & Management Science in the Age of Analytics*, pages 73–100. INFORMS, 2019.

[92] V. Marda and S. Narayan. Data in new delhi's predictive policing system. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 317–324, 2020.

[93] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[94] A. K. Menon and R. C. Williamson. The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*, pages 107–118. PMLR, 2018.

[95] Microsoft and Contributors. Fairlearn. `https://fairlearn.org/`, 2019.

[96] D. Mukherjee, M. Yurochkin, M. Banerjee, and Y. Sun. Two simple ways to learn individual fairness metrics from data. In *International Conference on Machine Learning*, pages 7097–7107. PMLR, 2020.

[97] G. Nakamura, B. E. Soares, V. D. Pillar, J. A. F. Diniz-Filho, and L. Duarte. Three pathways to better recognize the expertise of global south researchers. *npj Biodiversity*, 2(1):17, 2023.

[98] Y. Nakao, L. Strappelli, S. Stumpf, A. Naseer, D. Regoli, and G. D. Gamba. Towards responsible ai: A design space exploration of human-centered artificial intelligence user interfaces to investigate fairness. *International Journal of Human–Computer Interaction*, pages 1–27, 2022.

[99] Y. Nakao, S. Stumpf, S. Ahmed, A. Naseer, and L. Strappelli. Toward involving end-users in interactive human-in-the-loop ai fairness. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 12(3):1–30, 2022.

[100] D. Pedreschi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *KDD*, pages 560–568. ACM, 2008.

[101] D. Pedreschi, S. Ruggieri, and F. Turini. *Measuring Discrimination in Socially-Sensitive Decision Records*, pages 581–592. SIAM, 2009.

[102] L. Perini and J. Davis. Unsupervised anomaly detection with rejection. In *NeurIPS*, 2023.

[103] P. P. Ponce. Direct and indirect discrimination applied to algorithmic systems: Reflections to brazil. *Computer Law & Security Review*, 48:105766, 2023.

[104] A. Pugnana, L. Perini, J. Davis, and S. Ruggieri. Deep neural network benchmarks for selective classification. *arXiv preprint arXiv:2401.12708*, 2024.

[105] A. Pugnana and S. Ruggieri. AUC-based selective classification. In *AISTATS*, volume 206, pages 2494–2514. PMLR, 2023.

[106] A. Pugnana and S. Ruggieri. A model-agnostic heuristics for selective classification. In *AAAI*, pages 9461–9469. AAAI Press, 2023.

[107] E. M. Queiroga, M. F. Batista Machado, V. R. Paragarino, T. T. Primo, and C. Cechinel. Early prediction of at-risk students in secondary education: A countrywide k-12 learning analytics initiative in uruguay. *Information*, 13(9):401, 2022.

[108] A. Rhea, K. Markey, L. D'Arinzo, H. Schellmann, M. Sloane, P. Squires, and J. Stoyanovich. Resume format, linkedin urls and other unexpected influences on ai personality prediction in hiring: results of an audit. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 572–587, 2022.

[109] B. Richardson, J. Garcia-Gathright, S. F. Way, J. Thom, and H. Cramer. Towards fairness in practice: A practitioner-oriented rubric for evaluating fair ml toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.

[110] B. Richardson and J. E. Gilbert. A framework for fairness: A systematic review of existing fair ai solutions. *arXiv preprint arXiv:2112.05700*, 2021.

[111] J. Rood. Nyc local law 144- brief overview. `https://proceptual.com/2023/01/23/quick-takeaways-from-the-dcwp-rules-hearing-on-aedts-nyc-local-law-144/`, 2023.

[112] M. E. Rose and J. R. Kitchin. pybliometrics: Scriptable bibliometrics using a python interface to scopus. *SoftwareX*, 10:100263, 2019.

[113] B. Ruf and M. Detyniecki. Towards the right kind of fairness in ai. *arXiv preprint arXiv:2102.08453*, 2021.

[114] S. Ruggieri, D. Pedreschi, and F. Turini. Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(2):1–40, 2010.

[115] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.

[116] M. Sameki. Responsible ai dashboard: A one-stop shop for operationalizing responsible ai in practice. `https://techcommunity.microsoft.com/t5/azure-ai-blog/responsible-ai-dashboard-a-one-stop-shop-for-operationalizing/ba-p/3030944`, 2021.

[117] N. Schreuder and E. Chzhen. Classification with abstention but without disparities. In *UAI*, volume 161 of *Proceedings of Machine Learning Research*, pages 1227–1236. AUAI Press, 2021.

[118] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 59–68, 2019.

[119] A. Shah, Y. Bu, J. K. Lee, S. Das, R. Panda, P. Sattigeri, and G. W. Wornell. Selective regression under fairness criteria. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 19598–19615. PMLR, 2022.

[120] S. Shankar, R. Garcia, J. M. Hellerstein, and A. G. Parameswaran. Operationalizing machine learning: An interview study. *arXiv preprint arXiv:2209.09125*, 2022.

[121] H. Suresh and J. V. Guttag. A framework for understanding unintended consequences of machine learning. *ArXiv*, abs/1901.10002, 2019.

[122] Tensorflow. Tensorflow fairness indicators. `https://github.com/tensorflow/fairness-indicators`, 2020.

[123] The European Commission. The eu artificial intelligence act - article 14, 2023. `https://artificialintelligenceact.com/title-iii/chapter-2/article-14/`.

[124] M. van Bekkum and F. Z. Borgesius. Using sensitive data to prevent discrimination by artificial intelligence: Does the gdpr need a new exception? *Computer Law & Security Review*, 48:105770, 2023.

[125] M. Veale and R. Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2):2053951717743530, 2017.

[126] M. Veale, M. Van Kleek, and R. Binns. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14, 2018.

[127] S. Verma and J. Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE, 2018.

[128] S. Wachter, B. Mittelstadt, and C. Russell. Bias preservation in machine learning: the legality of fairness metrics under eu non-discrimination law. *W. Va. L. Rev.*, 123:735, 2020.

[129] S. Wachter, B. Mittelstadt, and C. Russell. Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai. *Computer Law & Security Review*, 41:105567, 2021.

[130] A. Wang, V. V. Ramaswamy, and O. Russakovsky. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, page 336–349, 2022.

[131] H. Wang, N. Grgic-Hlaca, P. Lahoti, K. P. Gummadi, and A. Weller. An empirical study on learning fairness metrics for compas data with human supervision. *arXiv preprint arXiv:1910.10255*, 2019.

[132] N. Wang. "black box justice": Robot judges and ai-based judgment processes in china's court system. In *2020 IEEE International Symposium on Technology and Society (ISTAS)*, pages 58–65. IEEE, 2020.

[133] Q. Wang, Z. Xu, Z. Chen, Y. Wang, S. Liu, and H. Qu. Visual analysis of discrimination in machine learning. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1470–1480, 2020.

[134] X. Wang and S. Yiu. Classification with rejection: Scaling generative classifiers with supervised deep infomax. In *IJCAI*, pages 2980–2986. ijcai.org, 2020.

[135] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019.

[136] S. C. Wheeler and R. E. Petty. The effects of stereotype activation on behavior: a review of possible mechanisms. *Psychological bulletin*, 127(6):797, 2001.

[137] M. Wick, s. panda, and J.-B. Tristan. Unlocking fairness: a trade-off revisited. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[138] M. Wick, J.-B. Tristan, et al. Unlocking fairness: a trade-off revisited. *Advances in neural information processing systems*, 32, 2019.

[139] J. N. Yan, Z. Gu, H. Lin, and J. M. Rzeszotarski. Silva: Interactively assessing machine learning fairness using causality. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pages 1–13, 2020.

[140] M. Yurrita, D. Murray-Rust, A. Balayn, and A. Bozzon. Towards a multi-stakeholder value-based assessment framework for algorithmic systems. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 535–563, 2022.

[141] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970. PMLR, 2017.

[142] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

[143] L. Zhang and X. Wu. Anti-discrimination learning: a causal modeling-based framework. *International Journal of Data Science and Analytics*, 4(1):1–16, 2017.

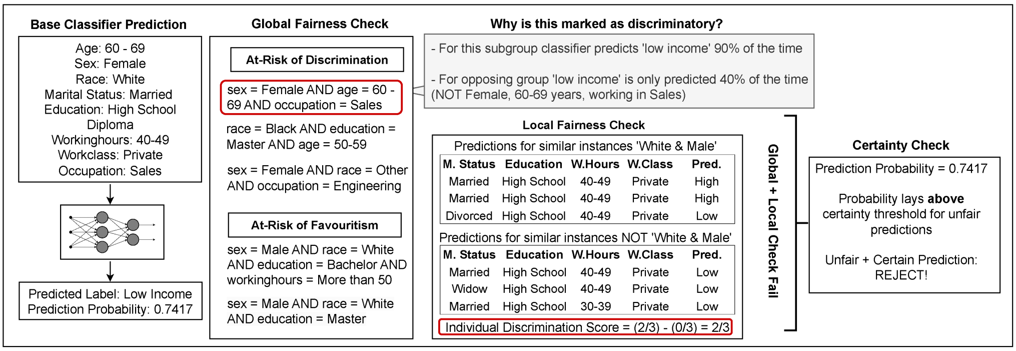
[144] L. Zhang, Y. Wu, and X. Wu. Situation testing-based discrimination discovery: A causal inference approach. In *IJCAI*, volume 16, pages 2718–2724, 2016.

Appendix A

Appendix Chapter 3

A.1 Illustrative Example of IFAC's Rejection Process

In Figure A.1 we see how our selective classification model IFAC behaves on one instance **x** of ACSIncome. In this example, a base classifier predicts that a **x** has a low income with a probability of 74.17%. To decide whether to keep this original prediction, IFAC starts by analysing if the prediction falls under any global patterns of unfairness it has recorded. In this case, the instance falls under the group of women, working in Sales aged between 60 and 69, that is marked as potentially discriminated. The reason why it is marked as such is that on a separate dataset, the ratio of negative prediction labels for this subgroup is much lower when the sensitive part describing this subgroup (in this case their sex) is negated. To illustrate: on this separate dataset the base-classifier predicted a negative decision label 90% of the time for the group women, working in Sales and aged between 60 and 69, as opposed to 40% for the same group of *non-female* instances. Given this high difference, the first global fairness check has failed, and the rejector proceeds with an individual fairness analysis. Here it makes use of the Situation Testing algorithm, and compares the positive label ratios of **x**'s most similar instances from the reference group (i.e. white men), with the positive label ratios of **x**'s most similar instances from the non-reference group. In doing so, it can make a more fine-grained fairness analysis, and not just assess the classifiers' behaviour on the group of people working in Sales and aged between 60 and 69; but also take into account other features, like peoples' education level or marital status. We observe here that even if individuals are similar regarding all legally grounded features, their sensitive characteristics still influence the ratio of positive decision labels, which is 2/3rd for our reference group white men and 0 for our non-reference group. Because this difference is quite large the local fairness test fails and the overall prediction is deemed as unfair. To then decide whether to perform a fairness intervention or reject the prediction, the rejector checks if the prediction probability of 74.17% falls above $t\_unfair\_certain$. In this case, it does, meaning that our prediction is unfair but certain. Hence, the rejector rejects the original low-income prediction. As a next step, this rejection and the explanation behind why the original prediction was considered unfair can be passed on to a human decision-maker. This person can use their domain knowledge as well as the explanation behind the rejection, to form a new decision for the instance in question. For instance, they may review the instances that were used for the similarity analysis in the individual fairness check, and determine if these instances were similar enough to the instance in


Base Classifier Prediction
Age: 60 - 69
Sex: Female
Race: White
Marital Status: Married
Education: High School Diploma
Workinghours: 40-49
Workclass: Private
Occupation: Sales
Predicted Label: Low Income
Prediction Probability: 0.7417
Global Fairness Check
At-Risk of Discrimination
sex = Female AND age = 60 - 69 AND occupation = Sales
race = Black AND education = Master AND age = 50-59
sex = Female AND race = Other AND occupation = Engineering
At-Risk of Favouritism
sex = Male AND race = White AND education = Bachelor AND workinghours = More than 50
sex = Male AND race = White AND education = Master
Why is this marked as discriminatory?
- For this subgroup classifier predicts 'low income' 90% of the time
- For opposing group 'low income' is only predicted 40% of the time (NOT Female, 60-69 years, working in Sales)
Local Fairness Check
Predictions for similar instances 'White & Male'
Predictions for similar instances NOT 'White & Male'
Individual Discrimination Score = (2/3) - (0/3) = 2/3
Global + Local Check Fail
Certainty Check
Prediction Probability = 0.7417
Probability lays above certainty threshold for unfair predictions
Unfair + Certain Prediction: REJECT!

Predictions for similar instances 'White & Male'

M. Status	Education	W.Hours	W.Class	Pred.
Married	High School	40-49	Private	High
Married	High School	40-49	Private	High
Divorced	High School	40-49	Private	Low

Predictions for similar instances NOT 'White & Male'

M. Status	Education	W.Hours	W.Class	Pred.
Married	High School	40-49	Private	Low
Widow	High School	40-49	Private	Low
Married	High School	30-39	Private	Low

Figure A.1: An illustrative example of how a low-income prediction for a woman from ACSIncome is deemed as discriminatory and subsequently rejected by our model

question to draw discrimination conclusions from. Further, the list of subgroups that the classifier behaves favourably/discriminatory on can serve to increase an expert's general understanding of the base classifier, and may be even adapted by them to incorporate their domain knowledge.

A.2 Proof: Setting slift threshold

In our methodology we select the discriminatory association rules used by IFAC, by checking for which of the rules the following property holds:

$$conf_{\mathbf{X}}((A,B) \rightarrow Y_v) - slift_{\mathbf{X}}((A,B) \rightarrow Y_v) < 0.5 \tag{A.1}$$

Which in the context of binary classification is true *iff*:

$$conf_{\mathbf{X}}((\neg A,B) \rightarrow Y_v) < conf_{\mathbf{X}}((\neg A,B) \rightarrow \neg Y_v) \tag{A.2}$$

Intuitively, this means that we only select the subgroups $\{A,B\}$ for which negating the sensitive part of the group ($\{\neg A,B\}$) yields a higher confidence for value Y_v w.r.t. the other value $\neg Y_v$.

Proof. Recalling the definition of $conf_{\mathbf{X}}((A,B) \rightarrow Y_v)$ as $P(Y_v|(A,B))$ we have that:

$$\begin{aligned} P(Y_v|(A,B)) - slift_{\mathbf{X}}((A,B) \rightarrow Y_v) &< 0.5 \\ P(Y_v|(A,B)) - (P(Y_v|(A,B)) - P(Y_v|(\neg A,B))) &< 0.5 \\ P(Y_v|(\neg A,B)) &< 0.5 \\ 2P(Y_v|(\neg A,B)) &< 1 \end{aligned} \tag{A.3}$$

For binary classification we can write $1 = P(Y_v|(\neg A, B)) + P(\neg Y_v|(\neg A, B))$ which yields:

$$\begin{aligned} 2P(Y_v|(\neg A, B)) &< P(Y_v|(\neg A, B)) + P(\neg Y_v|(\neg A, B)) \\ P(Y_v|(\neg A, B)) &< P(\neg Y_v|(\neg A, B)) \qquad \text{(A.4)} \\ conf_{\mathbf{X}}((\neg A, B) \rightarrow Y_v) &< conf_{\mathbf{X}}((\neg A, B) \rightarrow \neg Y_v) \end{aligned}$$

□

A.3 Full Fairness Results

In Table A.1 and A.2 we display the full fairness results for ACSIncome and Wisconsin-Recidivism for each classifier-methdology combination.

Table A.1: Full Fairness Results Income Prediction

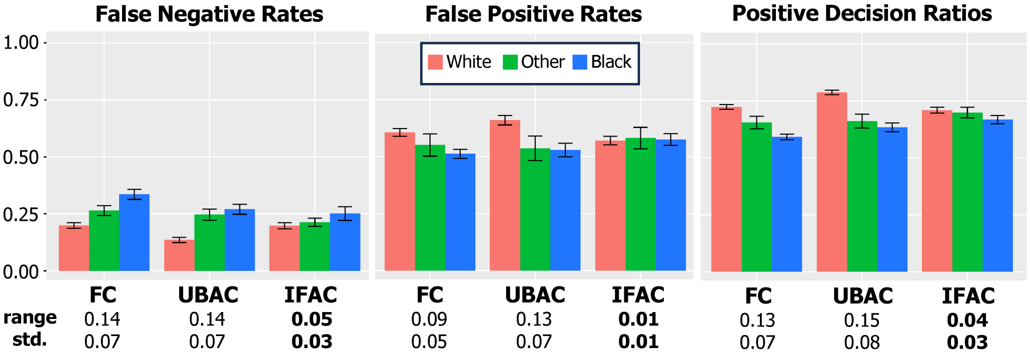
			M. Wh.	**F. Wh.**	**M. Bl.**	**F. Bl.**	**M. Oth.**	**F. Oth.**	**Range**	**Std.**
RF	FNR	FC	.33±.03	.57±.03	.57±.09	.60±.11	.44±.18	.59±.22	.27	.11
		UBAC	.26±.03	.54±.04	.61±.11	.67±.10	.30±.18	.54±.26	.40	.17
		IFAC	.37±.04	.44±.06	.57±.08	.49±.11	.41±.17	.52±.25	**.20**	**.08**
	FPR	FC	.24±.03	.10±.01	.12±.04	.05±.01	.08±.07	.05±.05	.19	.07
		UBAC	.20±.03	.06±.01	.07±.03	.02±.01	.07±.08	.03±.04	.18	.07
		IFAC	.18±.03	.11±.01	.10±.04	.04±.02	.08±.07	.05±.05	**.14**	**.05**
	Pos. Ratio	FC	.43±.02	.17±.01	.17±.03	.09±.01	.18±.07	.13±.07	.34	.12
		UBAC	.43±.03	.13±.01	.12±.03	.05±.02	.16±.07	.10±.07	.38	.13
		IFAC	.36±.02	.20±.01	.16±.03	.09±.02	.17±.08	.15±.07	**.27**	**.09**
NN	FNR	FC	.34±.03	.52±.04	.60±.08	.69±.09	.40±.22	.56±.22	.35	.13
		UBAC	.24±.04	.56±.06	.63±.09	.75±.10	.38±.22	.42±.26	.50	.18
		IFAC	.35±.04	.47±.07	.60±.08	.60±.14	.38±.22	.44±.29	**.25**	**.11**
	FPR	FC	.19±.02	.06±.01	.07±.03	.03±.01	.04±.04	.07±.04	.16	.06
		UBAC	.15±.02	.03±.01	.04±.03	.01±.01	.02±.03	.03±.04	.13	.05
		IFAC	.13±.01	.06±.01	.06±.03	.03±.02	.02±.03	.07±.04	**.11**	**.04**
	Pos. Ratio	FC	.40±.02	.15±.01	.14±.03	.07±.01	.15±.05	.16±.05	.34	.11
		UBAC	.40±.02	.09±.01	.10±.03	.03±.01	.12±.06	.11±.06	.37	.13
		IFAC	.33±.02	.15±.01	.12±.03	.07±.01	.12±.06	.15±.05	**.27**	**.09**
XGB	FNR	FC	.29±.03	.57±.05	.57±.09	.62±.07	.36±.14	.52±.25	.33	.13
		UBAC	.20±.03	.62±.07	.65±.12	.80±.08	.16±.16	.43±.28	.65	.26
		IFAC	.33±.03	.47±.06	.61±.10	.62±.11	.38±.15	.40±.26	**.29**	**.12**
	FPR	FC	.19±.02	.05±.01	.07±.02	.04±.01	.08±.05	.03±.04	.16	.06
		UBAC	.14±.02	.02±.01	.03±.02	.02±.01	.03±.04	.02±.02	.12	.05
		IFAC	.11±.02	.06±.01	.06±.01	.03±.01	.06±.06	.02±.04	**.09**	**.03**
	Pos. Ratio	FC	.42±.02	.13±.01	.14±.02	.08±.02	.19±.06	.13±.07	.34	.12
		UBAC	.41±.02	.08±.02	.09±.03	.04±.01	.15±.07	.10±.06	.38	.14
		IFAC	.32±.02	.15±.01	.12±.03	.06±.02	.16±.06	.13±.07	**.27**	**.09**

Table A.2: Full Fairness Results Recidivism Prediction

			White	**Black**	**Other**	Range	Std.
RF	FNR	BC	.20 ± .01	.34 ± .02	.26 ± .02	.14	.07
		USC	.14 ± .01	.27 ± .02	.25 ± .02	.13	.07
		FSC	.14 ± .01	.24 ± .02	.24 ± .02	.10	.05
	FPR	BC	.61 ± .02	.51 ± .02	.55 ± .05	.09	.05
		UBAC	.66 ± .02	.53 ± .03	.54 ± .05	.13	.07
		IFAC	.64 ± .02	.56 ± .03	.56 ± .06	.08	.05
	Pos. Ratio	FC	.72 ± .01	.59 ± .01	.65 ± .03	.13	.07
		UBAC	.79 ± .01	.63 ± .02	.66 ± .03	.15	.08
		IFAC	.77 ± .01	.66 ± .02	.67 ± .03	.11	.06
NN	FNR	FC	.22 ± .01	.38 ± .02	.30 ± .02	.17	.08
		UBAC	.20 ± .01	.34 ± .02	.27 ± .02	.14	.07
		IFAC	.20 ± .01	.33 ± .02	.26 ± .02	.13	.06
	FPR	FC	.58 ± .02	.44 ± .02	.51 ± .06	.14	.07
		UBAC	.56 ± .02	.42 ± .02	.50 ± .05	.14	.07
		IFAC	.55 ± .02	.43 ± .02	.51 ± .05	.12	.06
	Pos. Ratio	BC	.70 ± .01	.53 ± .01	.62 ± .03	.17	.09
		UBAC	.71 ± .01	.55 ± .01	.63 ± .03	.16	.08
		IFAC	.70 ± .01	.56 ± .01	.64 ± .03	.14	.07
XGB	FNR	FC	.20 ± .01	.33 ± .03	.26 ± .02	.14	.07
		UBAC	.14 ± .01	.28 ± .02	.23 ± .02	.14	.07
		IFAC	.14 ± .01	.28 ± .02	.23 ± .02	.14	.07
	FPR	FC	.60 ± .01	.46 ± .03	.57 ± .03	.15	.07
		UBAC	.65 ± .02	.47 ± .04	.51 ± .03	.18	.09
		IFAC	.64 ± .02	.46 ± .04	.51 ± .03	.18	.09
	Pos. Ratio	BC	.72 ± .01	.56 ± .02	.67 ± .02	.16	.08
		UBAC	.78 ± .01	.60 ± .02	.66 ± .02	.18	.09
		IFAC	.78 ± .01	.60 ± .02	.67 ± .02	.18	.09

A.4 WisconsinRecidivism Results with Less Strict Unfairness Selection

In Figure A.2 we see the results of a Random Forest classifier combined with the different abstention methods on WisconsinRecidivism. For the local fairness check as executed with Situation Testing we now set the threshold t to 0.0. Intuitively this means, that regardless of the local fairness results any instance falling under a global pattern of discrimination will be considered as unfair (the situation testing results can still be used as extra information for a human reviewer). We see here that with this less strict unfairness selection, IFAC reduces FNR, FPR and PDR differences across demographics more than when using $t = 0.3$.



	FC	UBAC	IFAC	FC	UBAC	IFAC	FC	UBAC	IFAC
range	0.14	0.14	**0.05**	0.09	0.13	**0.01**	0.13	0.15	**0.04**
std.	0.07	0.07	**0.03**	0.05	0.07	**0.01**	0.07	0.08	**0.03**

Figure A.2: Caption

A.5 Effects of c and w_u

In Figure A.3 we display the effects of both the coverage parameter c and the unfair-reject-weight w_u on the accuracy as well as the fairness of our abstention method IFAC. We compare the results with a regular uncertainty based abstaining classifier (UBAC) and a full covage (FC) one.

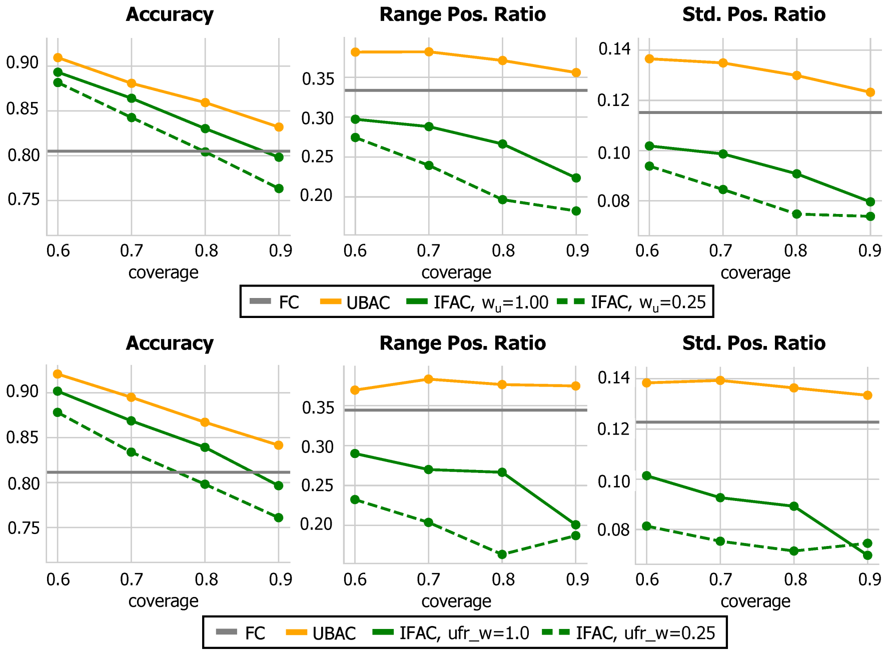


Figure A.3: ACSIncome effect of different values for c and w_u on abstention methods combined with Neural Network (above) and XGBoost

Appendix B

Appendix Chapter 4

B.1 Datasheet for Dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to provide a new benchmarking tool for fair Machine Learning algorithms. Different than other datasets typically used for evaluating fair ML, our data contains a fair and biased version of its decision label. Using this, we can check the effectiveness of a fair ML intervention, by checking how well it can predict the fair labels after being trained on the biased ones.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Daphne Lenders and Toon Calders were responsible for collecting the biased labels of this dataset. Both are affiliated with the ADReM Data Lab of the University of Antwerp. The rest of the data (including its fair label) was based on an already existing dataset, which is publicly available online[1].

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

The creation of the dataset was funded by the University of Antwerp - Research Excellence Center DigiTax.

Any other comments? /

[1] `https://www.kaggle.com/datasets/uciml/student-alcohol-consumption` (license CC0)

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Each instance in this dataset represents a high school student, following either a Portuguese or Maths course. There is a variety of information available for each student, including, e.g., information about their studytime, their school absences and their free time behaviour. Special variable of interests are students' sex, their performance on an exam (pass vs. fail) and their predicted performance on that exam. The predicted performance was based on an experiment, where students were presented with some information about the students, based on which they had to make grade predictions. Comparing the predicted exam performance with the actual exam performance we observe clear bias against boys.

How many instances are there in total (of each type, if appropriate)?

Our dataset consists of a total of 856 instances. Note, that the two dataset entries might relate to the same student, whereas one entry corresponds to the student's performance in a Maths course and the other to the performance in a Portuguese course.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The instances in our dataset are sampled from a larger dataset that is publicly available. In sampling from this data, we excluded all students whose grade for the last exam of the course was 0. Further, we randomly sampled 428 from the 430 male instances, and 428 from the 560 female instances. These steps were taken so that in the collection of our biased labels, we could present each participant with four male and four female student profiles, which the participants had to make grade predictions for.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance consists of several features, that are either categorical or numerical in nature.

Is there a label or target associated with each instance? If so, please provide a description.

While other features of the data could potentially be used as a target as well, our intended target label is whether each student passed or failed the third exam of the course they were following. Our dataset consists of both a biased and a fair version of this label. The fair version was obtained, by checking whether a student's grade for the last exam was ≥ 10. There are multiple ways in which the biased label can be obtained and all are based on a human experiment where participants ranked eight student profiles according to their expected performance and predicted their grade for the exam. The first way to obtain the biased label is by checking whether the predicted grade for the exam is ≥ 10. The second way is to look at the ranking position each student was assigned to: the passing-labels of most highly ranked instances were always changed to "true", while they were changed to "false" for the two lowest rated instances.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No information is missing.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

Relationships between individual instances are not made explicit.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

There are no recommended data splits.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Since our dataset was partly based on an already existing dataset, and partly based on our own experiment there are two types of noise in the data. The first relates to the noise that was already present in the original data, which consists of information of high school students and their exams for a course. A lot of information of the high school students was self-recorded, like, e.g., information on their drinking behaviour or their studytime. Thus it is questionable to which extent the students were truthful in reporting this information. Also, in the collection of this data, the sex of the students was treated as a binary variable. Thus, important information on non-binary gender identity may have been lost. Second, our biased labels also contain some noise, due to the fact that they were gathered through a human experiment. In this experiment 107 participants each made grade predictions for 8 student profiles. Because each participant might have different stereotypes and biases when estimating students' exam performance, the labels may not be very consistent. It should be noted that this noise was intentionally introduced, to reflect the complexity of real-life decision making and discriminatory biases.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

While the main part of our dataset can be used as it is, it is possible to extent it further using the original data it was based on. As this data was publicly available online (license CC0), we included a preprocessed version of it on our kaggle and github page. This data contains some information of the students, that we did not present when asking participants to make grade predictions for them. It is possible to link this information to our collected data, using the indices of both datasets. Because we make both the preprocessed original data and our collected data available online, we can guarantee that both will keep existing without any changes (except changes we might make ourselves).

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

No, this dataset does not contain any of such data.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

As mentioned, our dataset consists of both a fair and biased version of its decision label. Comparing the biased label with the fair one, we see that participants had discriminatory biases (mostly targeted against boys) when making grade predictions. These biases may be offensive. Also, “sex" is treated as a binary variable in our dataset, which was a direct consequence of the binary gender categories used in the original “Student Alcohol Consumption" dataset. While we recognize that this may be offensive to some people (especially those who identify with a non-binary gender category), we emphasize that this does not reflect our own beliefs about gender identities.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, the dataset relates to people.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

Yes, our datast contains information about students’ sex, and it is also possible to infer their age using the original data it was based on. In our dataset half of the students (i.e. 428 out of 856) are male and the other half are female. The ages of students range between 15 and 22. There are 169 students who are 15 years old, 230 who are 16 years, 230 who are 17 years, 172 who are 18 years, 41 who are 19 years and 15 students who are 20 years or older (9 students that are 20, 3 that are 21 and 2 that are 22).

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No, it is not possible to identify individuals from out dataset.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

The sex of the students might be considered as sensitive data. Further, in the original data there is some information about the parents of the students including their job and education level. Both might give some indication of the socio-economic status of the students. Finally, the data also contains information about the drinking behaviour of students. This information can especially be considered as sensitive, given that not all of the students are of legal drinking age (in Portugal, the country where the data was collected this age is 18+). We emphasize that we do not want to encourage illegal drinking, by distributing our dataset. Still, we also highlight that none of the students of which the data was collected have to fear unwanted consequences for their actions, as the data is completely anonymous and cannot be traced back to any individuals.

Any other comments? /

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The information about the students as well as their grades were gathered in a previous study by Cortez and Silva. The data was gathered from two high schools in Portugal, for more information on how they collected this data we refer to their paper[2]. We collected the biased labels for this dataset through a survey, where participants were asked to make grade predictions for students, based on short descriptions about them. In the paper corresponding to our dataset, we give a more detailed description of this survey, including the exact task set-up, the materials we used and information about participant recruitment procedures.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

[2] Cortez, P. & Silva, A. (2008). Using data mining to predict secondary school student performance. In A. Brito and J. Teixeira Eds., *Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008)*, pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

We used the survey platform Qualtrics[3] to set up our survey. Our survey was based on multiple pilot studies, were we tested its clarity and suitability.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

As previously mentioned the data we collected our biased labels for, was sampled from a slightly bigger dataset that was already publicly available online. We excluded all students with a grade equal to 0 from this dataset, and randomly sampled 428 male and 428 female instances from the rest of the data.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The collection of our data was based on voluntary participation. Further, we also put our survey on the Survey Exchange platforms SurveySwap and SurveyCircle[4]. Here participants who filled out our survey were rewarded with survey-responses for their own survey.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The biased labels were sampled from January until March 2022. This timeframe does not match the creation of the data associated with the instances: the information about each student was recorded in 2005 and 2006.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

To collect our biased labels we obtained ethical approval by the *Ethics Committee for the Social Sciences and Humanities* of the University of Antwerp, under reference number SHW_21_128. To start the review process we gave a detailed description of our experimental setup, possible risks involved in participation and the way in which we would store and process the collected data. We received a positive outcome for this review process.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, the dataset relates to people.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

[3] `https://www.qualtrics.com`

[4] `https://surveyswap.io/` and `https://www.surveycircle.com/en/surveys/`

We obtained the data about the students through an already existing dataset (this existing dataset was based on asking individuals directly for their information). The biased labels were also obtained directly, by specifically asking our participants to make grade predictions for the students.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Again, for details on the data collection of the original data (with the information about the students) we refer to the paper of Cortez and Silva. To collect the biased labels for this dataset, participants first had to fill out a consent form, before they could start the survey. Here it was also explained that the collected data would be made publicly available.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Yes, our participants had to consent to the collection and use of their data. The consent form can be seen when following our survey link: `https://uantwerpen.eu.qualtrics.com/jfe/form/SV_5gOFzeF3xtGSinI`

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

As the participants' data was collected anonymously, and survey responses could not be matched to individuals' identity, we did not provide participants with a mechanism to revoke their consent.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Because we are dealing with anonymous data, no formal analysis has been conducted.

Any other comments? /

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Before collecting our biased decision labels based on the original dataset, we applied some pre-processing steps on it. All steps are described below:

- **Parent's education** - This was a variable that was not part of the original dataset. Instead the dataset consisted of two variables, namely *Fedu* and *Medu* to respectively denote the father's and the mother's education level. We obtained our variable *Parent's education* by taking the maximum of the two.
- **Studytime** - Originally, this variable consists of four levels (*less than 2 hours* vs. *2-5 hours* vs. *5-10 hours* vs. *more than 10 hours*). Because there were little students with a studytime of longer than 10 hours, we decided to merge the latter two levels
- **Absences** Originally, this variable ranges from 0 to 93. Since high values for this variable were quite uncommon, we decided to bin all absences ≥ 7 into one level called *More than 6*
- **Freetime** - Originally, this variable ranged from 1 (very low) to 5 (very high). We binned this variable into three categories, where 1 & 2 are binned into level, and 4 & 5 are binned as well
- **Gooing out** - Again, this variable originally ranged between 1 and 5. We decided to bin the last two levels (4 & 5)
- **Alcohol Consumption** - The original dataset consisted of two variables denoting the student's alcohol consumption namely *Walc* (alcohol consumption in the weekend) and *Dalc* (alcohol consumption throughout the week). For our experiment we only showed the students' alcohol consumption in the weekend. This variable originally consisted of 5 levels, where we binned the latter two ones

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

Yes, a not-preprocessed version of the data is being provided on our kaggle page as well.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the (Python) code that we used to preprocess the data is available on github: `https://github.com/calathea21/settingUpBenchmarkCollection`

Any other comments? /

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

The dataset has been used for our own benchmarking experiment. Here we tested the effectiveness of several fairness interventions, by checking how well they can predict the fair labels of the dataset after being trained on the biased ones.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

No such repository is available.

What (other) tasks could the dataset be used for?

The dataset was created mostly for benchmarking studies, like the one previously described. However, there are some other interesting use cases for the fair Machine Learning community. It could for instance be interesting to use the data to better understand the dynamics behind discriminatory decision making, by checking how exactly the fair labels relate to the biased ones, and if there are some clear patterns in which discrimination/-favouritism occurs. This knowledge could then also be exploited to create better fairness interventions.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

Though the dataset can be a useful tool for benchmarking fair ML algorithms, users should be careful not to overgeneralize their findings to other decision tasks. A fairness intervention that performs well on our dataset, is not guaranteed to work well on others, and may not be as fair/accurate as intended. Also, users should be cautious with the fact that we treat “sex" as a binary variable in our dataset. If researchers start developing new fairness algorithms based on our data, they should take into account that our data does not provide information on the types of discrimination non-binary people might face.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Our dataset was made to test the effectiveness of fairness interventions targeting label bias. We say that our collected version of the decision labels (i.e. whether students pass an exam or not) contain label bias as they do not accurately reflect whether the students actually passed or not, and where instead the result of a biased decision process. Label bias is different than other forms of biases, like selection bias or historical bias. Hence, fairness interventions that specifically target this kind of bias, should not be benchmarked on our dataset.

Any other comments? /

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Our dataset is publicly available online (under license CC BY-SA 4.0[5]) meaning that any third party can access and use it, as long as they give appropriate credit, provide a link to the license and indicate if changes were made to the data.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

The dataset is distributed via kaggle: `https://www.kaggle.com/datasets/daphnelenders/performance-vs-predicted-performance`. It's DOI is: `10.34740/kaggle/dsv/3689065`

When will the dataset be distributed?

Our dataset is already publicly available on kaggle.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

Our dataset is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License. To view a copy of this license, visit `http://creativecommons.org/licenses/by-sa/4.0/` or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No restrictions has been imposed on the data.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No export controls or other regulatory restrictions apply to the dataset.

Any other comments? /

Maintenance

[5] `http://creativecommons.org/licenses/by-sa/4.0/`

Who will be supporting/hosting/maintaining the dataset?

Daphne Lenders will be responsible for supporting, hosting and maintaining the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Daphne Lenders can be contacted by her institutional email: daphne.lenders@uantwerpen.be

Is there an erratum? If so, please provide a link or other access point.

As for now there is no erratum.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

The dataset was based on an existing dataset and on a one-time human experiment. Unless new experiments are conducted, or mistakes in the current data are found, it is unlikely that the dataset will be updated.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

There are no limits on the retention of the data.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

If new versions of the dataset will be made available, all older versions will still be accessible through the kaggle website.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

No formal mechanism for contributing to our dataset consists yet, as for now it is unlikely to be expanded. If there are suggestions for extending or augmenting the data, it is possible to send an email to daphne.lenders@uantwerpen.be.

Any other comments? /

B.2 Proof-of-Concept Study

In this section we are going to give additional information about the study design of our proof-of-concept as well as its results. We set up this study as a precedent for our main experiment, where we test whether people have inherent biases against boys when judging their school performance and whether this bias can be triggered or amplified through stereotype activation. The study was approved by the Ethics Committee of the University of Antwerp, under reference number SHW_21_128.

B.2.1 Experimental Design

The experimental design of our proof-of-concept study is illustrated in Figure B.1. In the main task, we presented 8 student profiles to the participants, containing basic information about each student, for which participants had to make grade predictions. The main manipulation was that part of the participants were presented with a version of these profiles for which the sex was swapped. In other words, a profile that belongs to a female student is, in this condition, said to belong to a male student (and vice versa). By comparing the predicted grades across these two conditions, we could determine whether participants have inherent biases against boys when estimating students' performance. To see whether bias against boys can be triggered, our second manipulation in the experiment was whether participants were exposed to some form of stereotype activation prior to the prediction task. We included three types of stereotype activation, which we will describe in sections. Note, that the complete experimental design, including the type of materials shown to the participants, the task set-up, as well as the task description, were based on a total of three pilot studies. In these pilot studies we let 4-6 participants complete the most up-to-date version of the online survey, and based on their responses and feedback we iteratively improved the overall study design.

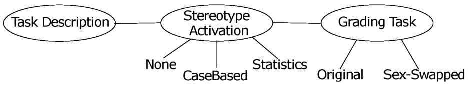
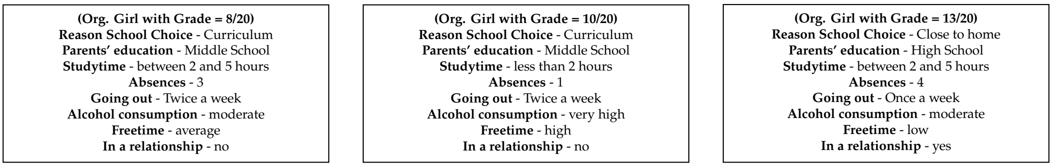


Figure B.1: The experimental setup to see if participants have inherent bias against boys or whether this bias can be triggered through stereotype activation

B.2.2 Materials

For the main task of the study, each participant was presented with the same eight student profiles extracted from the "Student Alcohol Consumption" dataset. Half of these profiles belong to male, and the other half to female students. While the original dataset contains more than 30 attributes to describe each student, we chose to only present eight of them per profile, to not overload participants with information. We chose attributes that had high variability between students and that could in legitimate or stereotypical ways be associated with school performance. In Figure B.2 we show some examples of the presented profiles, along with the original sex of each student and their obtained grade.



(Org. Girl with Grade = 8/20)
Reason School Choice - Curriculum
Parents' education - Middle School
Studytime - between 2 and 5 hours
Absences - 3
Going out - Twice a week
Alcohol consumption - moderate
Freetime - average
In a relationship - no

(a) Student Profile 1

(Org. Girl with Grade = 10/20)
Reason School Choice - Curriculum
Parents' education - Middle School
Studytime - less than 2 hours
Absences - 1
Going out - Twice a week
Alcohol consumption - very high
Freetime - high
In a relationship - no

(b) Student Profile 2

(Org. Girl with Grade = 13/20)
Reason School Choice - Close to home
Parents' education - High School
Studytime - between 2 and 5 hours
Absences - 4
Going out - Once a week
Alcohol consumption - moderate
Freetime - low
In a relationship - yes

(c) Student Profile 3

Figure B.2: Some of the profiles that participants had to make grade predictions for

As can be seen in this Figure, each student profile was presented in a tabular format. To convey the sex of each student, we randomly assigned each profile to one of four male/female names, depending on the students' sex in the original dataset and whether the profile was in the 'sex-swapped' condition or not. The four male and female names were chosen to represent common names in English speaking countries. In the experiment all eight profiles were presented on one page, where the order of presentation was randomized. On top of this page, participants were presented with a list of all student names followed by a blank field. They were asked to use a drag-and-drop interface to rank the students according to their expected performance. Additionally, they were prompted to enter specific grade predictions (ranging from 0 to 20) in the blank field next to each students' name.

Before the grading task, participants were exposed to one of three forms of stereotype activation:

1. **None** - Baseline condition in which no extra information is presented.
2. **CaseBased** - Here we presented participants with three student profiles along with the grades of the students. Two profiles belong to male students with low grades (5/20 and 10/20), while one belongs to a female student with a high grade (17/20).
3. **Statistics** - Here we presented a graph showing statistics about how some risk factors affect boys' chance to pass an exam more than they affect girls' passing chances. One presented risk factor was, e.g., having more than 6 school absences, which makes boys ~15% more likely to fail, while girls only ~4% more likely. All risk factors were chosen such that none of the presented profiles contained any of these risk factors.

As can be seen, both the CaseBased and Statistics condition contained stereotypical information against boys. We were interested to see whether presenting this information prior to the prediction task would differently affect the grade predictions for male and female students.

B.2.3 Participants

The participants were recruited through social media channels and the survey exchange platforms SurveySwap and SurveyCircle. To participate, a consent form needed to be filled out. All responses were completely anonymous, and after a quality controls[6], to

[6] see: `https://github.com/calathea21/analyzing_proof_of_concept`

filter out short responses and respondents who did not follow the survey instructions correctly, we were left with data of 157 participants.

In table B.1 we show how the participants were distributed over the different conditions of our experiment. In table B.2 the participant counts per gender and age category are shown. Figure B.3 gives an overview of the nationality of the participants.

Table B.1: Number of Participants Across the Conditions

Stereotype Act. Datatype	None	Case Based	Statistics
Original	N = 28	N = 28	N = 27
Sex-Swapped	N = 29	N = 30	N = 27

Table B.2: Number of participants by gender and by age

Gender	**Count**	**% of total**
Female	107	63.3%
Male	60	35.5%
Prefer not to say	2	1.2%

Age	**Count**	**% of total**
18 - 24	97	57.4%
25 - 34	63	37.3%
35 - 44	7	4.1%
65 - 74	1	0.6%
Prefer not to say	1	0.6%

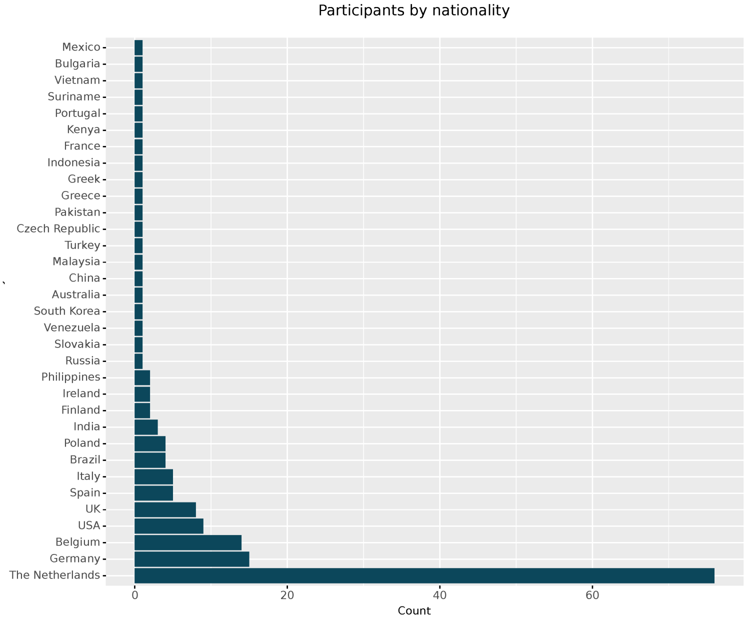


Figure B.3: Participants by Nationality

B.2.4 Data Analysis

On the grade predictions of all eight student profiles we applied the Align Rank Transform procedure with an ANOVA, a nonparametric factorial analysis technique introduced by Wobbrock et al[7]. As factors we used “Stereotype Activation" and “Sex" (i.e. whether a profile was said to belong to a male or female student), while the “Predicted Grade" for each student profile was used as the dependent variable. For post-hoc tests for pairwise comparison, we used Tukey correction. Note, that we chose to not apply multiple comparison correction for the different ANOVAs. Using a significance cut-off value of 0.05 and conducting eight ANOVAs each testing three hypothesis, this results into 1.2 expected false positives (24*0.05). Given that our proof-of-concept study is more of preliminary nature and we also did not want to compromise on the statistical power of our analysis, we deemed this as acceptable.

B.2.5 Illustration of Most Important Results

In this section we are going to demonstrate the different types of effect we found on three student profiles. We refer to Appendix B.2.6 for the statistical test results on all eight

[7] Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011, May). The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 143-146).

profiles. The results for this section are visualized in Figure B.4, where Profile 1, 2 and 3 correspond to the profile descriptions in Figure B.2.

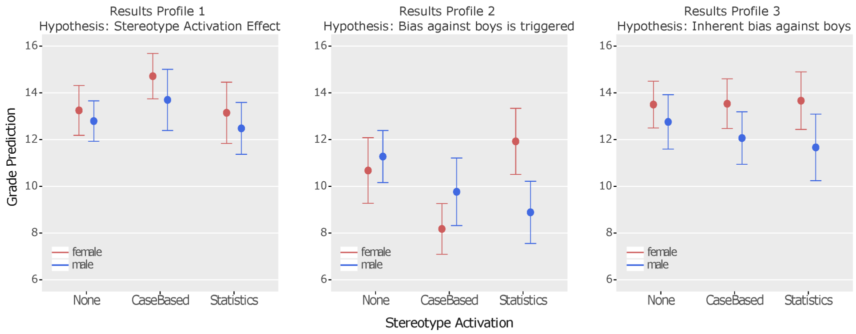


Figure B.4: Participants' grade predictions can depend on the stereotype activation they were exposed to and on the presented sex of the student profile (as well as on the interaction between both)

For student profile 3 (right graph of Figure B.4), we found a significant effect of “sex" on grade prediction. Averaged across all stereotype activation conditions, higher grades were predicted for this profile if it was tied to a girl's rather than a boy's name (F(1, 163) = 5.255, p = 0.023), with a mean difference of 1.40. This finding confirms our hypothesis that in some cases humans have inherent stereotypes against male students and expect them to perform less well in high school than girls.

On the grade predictions on student profile 1 (left graph of Figure B.4) we found a significant effect of “stereotype activation" on grade predictions (F(2, 163) = 4.031, p = 0.020). Averaging over the grades assigned to the male and female version of the profile, participants who were exposed to the “CaseBased" condition gave significantly higher grades than participants exposed the “Statistics" stereotype activation (with a mean difference of 1.392). A similar effect was found in student profile 2 (middle graph of Figure B.4, F(2, 163) = 5.157, p = 0.007); here participants presented with the CaseBased condition predicted significantly lower grades than participants in the other conditions. Even though we did not make any hypotheses about the effects of stereotype activation alone, these observations were interesting to see. We hypothesise that an anchoring effect occurred, where information about the grades of other students (as presented in the CaseBased condition) influenced participants' subsequent grade predictions [8]. Differently than expected, this effect does not influence the grade predictions for male and female students differently. This might be the case because in the CaseBased condition only three student profiles (2 male, 1 female) were shown. The presented sex-differences in their grades might have been too subtle, for stereotypes to be activated.

For student profile 2, we observed, next to the significant effect of “stereotype activation" alone, a significant interaction effect of “sex" and “stereotype activation" on the grade predictions. In other words, participants assign different grades to the male and female version of this profile, depending on which stereotype activation was presented (F(2, 163) = 8.402 p = 0.001). In this case, the difference in female and male grade for

[8] Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The journal of socio-economics*, 40(1), 35-42.

participants under the “Statistics" condition (diff$_{\text{female - male}}$ = 3.037) is higher than under the “CaseBased" (diff$_{\text{female - male}}$ = -1.588) and “None" conditions (diff$_{\text{female - male}}$ = -0.597). This confirms our hypothesis that biases against boys are not always inherent but can be triggered, in particular through the “Statistics" stereotype activation.

One more general finding of our study is that significant effects of “stereotype activation" or “sex" (or their interaction), were only found on certain profiles. In particular, we observed that no effects occurred on “stereotypically good" profiles, of students with e.g. high amount of study time or low alcohol consumption (see Appendix, profile 4 and 8). Even though we did not go into a deeper analysis of this, it confirms our hypothesis that the occurrence of bias does not only depend on the sex of the students, but also on their other, more complex characteristics. While a more elaborate study is needed to generalize our findings to real-life human behaviour, our results show that the experimental setup of our study is appropriate to elicit interesting biases in human decision makers. As some of these biases are discriminatory, we deemed the setup as useful for our main study.

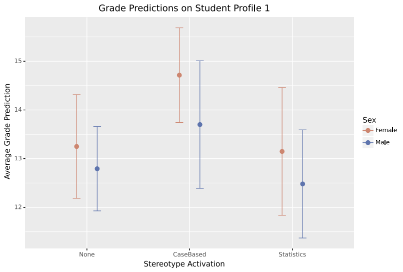
B.2.6 All Statistical Results

Now that we have illustrated the type of main- and interaction effects we found in our proof-of-concept study, we will show the statistical results on all the student profiles.

B.2.6.1 Profile 1

(Org. Girl with Grade = 8/20)
Reason School Choice - Curriculum
Parents’ education - Middle School
Studytime - between 2 and 5 hours
Absences - 3
Going out - Twice a week
Alcohol consumption - moderate
Freetime - average
In a relationship - no

(a) Student Profile 1



(b) Average grade predictions

Figure B.5: Results on Student Profile 1

Table B.3: Results Statistical Test Profile 1

	Sum of Squares	df	Mean Square	F	p
Datatype	6217	1	6217	2.570	0.111
Stereotype Activation	18868	2	9434	4.0309	0.020
Datatype * Stereotype Activation	14.9	2	7.47	0.0030	0.997

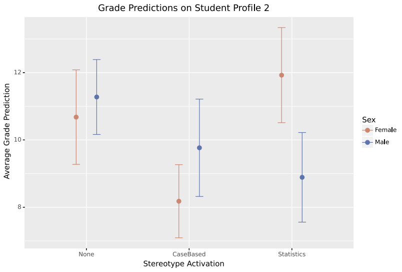
Table B.4: Results Post Hoc Test Profile 1

Comparison							
Stereotype Activation		**Stereotype Activation**	**Mean Difference**	**SE**	**df**	**t**	**p_{tukey}**
CaseBased	-	None	21.30	9.03	163	2.360	0.051
CaseBased	-	Statistics	23.17	9.15	163	2.531	0.033
None	-	Statistics	1.87	9.19	163	0.203	0.978

B.2.6.2 Profile 2

(Org. Girl with Grade = 10/20)
Reason School Choice - Curriculum
Parents' education - Middle School
Studytime - less than 2 hours
Absences - 1
Going out - Twice a week
Alcohol consumption - very high
Freetime - high
In a relationship - no

(a) Student Profile 2



(b) Average grade predictions

Figure B.6: Results on Student Profile 2

Table B.5: Results Statistical Test Profile 2

	Sum of Squares	**df**	**Mean Square**	**F**	**p**
Datatype	637	1	637	0.260	0.611
Stereotype Activation	23819.80	2	11909.90	5.157	0.007
Datatype * Stereotype Activation	37551.4	2	18775.7	8.4024	0.001

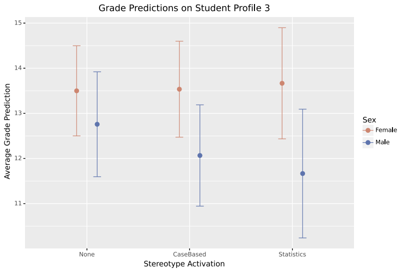
Table B.6: Results Post Hoc Test Profile 2

Comparison							
Stereotype Activation		**Stereotype Activation**	**Mean Difference**	**SE**	**df**	**t**	**p_{tukey}**
CaseBased	-	None	-26.95	8.97	163	-3.006	0.009
CaseBased	-	Statistics	-22.33	9.09	163	-2.456	0.040
None	-	Statistics	4.63	9.13	163	0.507	0.868

B.2.6.3 Profile 3

(Org. Girl with Grade = 13/20)
Reason School Choice - Close to home
Parents' education - High School
Studytime - between 2 and 5 hours
Absences - 4
Going out - Once a week
Alcohol consumption - moderate
Freetime - low
In a relationship - yes

(a) Student Profile 3



(b) Average grade predictions

Figure B.7: Results on Student Profile 3

Table B.7: Results Statistical Test Profile 3

	Sum of Squares	df	Mean Square	F	p
Datatype	12536	1	12535.5	5.255	0.023
Stereotype Activation	1527.5	2	763.7	0.311	0.733
Datatype * Stereotype Activation	3242.391	2	1621.196	0.664	0.516

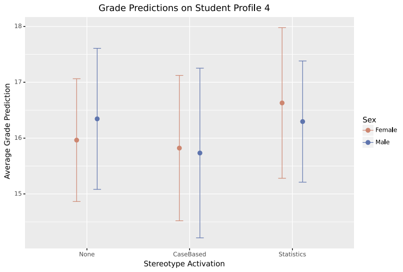
Table B.8: Results Post Hoc Test

Comparison							
Datatype		**Datatype**	**Mean Difference**	**SE**	**df**	**t**	**p_{tukey}**
Sex-Swapped (Male)	-	Original (Female)	-17.2	7.52	163	-2.29	0.023

B.2.6.4 Profile 4

(Org. Girl with Grade = 15/20)
Reason School Choice - Reputation
Parents' education - University
Studytime - more than 5 hours
Absences - 0
Going out - Twice a week
Alcohol consumption - high
Freetime - high
In a relationship - yes

(a) Student Profile 4



(b) Average grade predictions

Figure B.8: Results on Student Profile 4

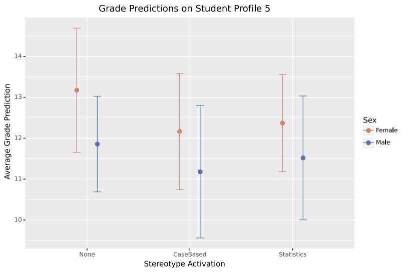
Table B.9: Results Statistical Test Profile 4

	Sum of Squares	df	Mean Square	F	p
Datatype	128	1	128	0.0523	0.819
Stereotype Activation	2658.2	2	1329.1	0.5438	0.582
Datatype * Stereotype Activation	4338	2	2169	0.894	0.411

B.2.6.5 Profile 5

(Org. Boy with Grade = 8/20)
Reason School Choice - Unknown
Parents' education - Middle School
Studytime - between 2 and 5 hours
Absences - 2
Going out - Once a week
Alcohol consumption - very high
Freetime - low
In a relationship - yes

(a) Student Profile 5



(b) Average grade predictions

Figure B.9: Results on Student Profile 5

Table B.10: Results Statistical Test Profile 5

	Sum of Squares	df	Mean Square	F	p
Datatype	9711.7	1	9711.71	4.045	0.046
Stereotype Activation	3880.80	2	1940.40	0.797	0.453
Datatype * Stereotype Activation	482.29	2	241.15	0.098	0.907

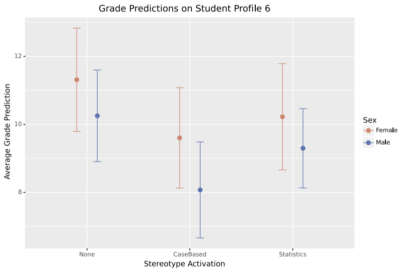
Table B.11: Results Post Hoc Test

Comparison							
Datatype		Datatype	Mean Difference	SE	df	t	p_{tukey}
Sex-Swapped (Male)	-	Original (Female)	15.2	7.54	163	2.01	0.046

B.2.6.6 Profile 6

(Org. Boy with Grade = 10/20)
Reason School Choice - Curriculum
Parents' education - Middle School
Studytime - less than 2 hours
Absences - 4
Going out - Twice a week
Alcohol consumption - high
Freetime - high
In a relationship - no

(a) Student Profile 6



(b) Average grade predictions

Figure B.10: Results on Student Profile 6

Table B.12: Results Statistical Test Profile 6

	Sum of Squares	df	Mean Square	F	p
Datatype	9710	1	9710	4.041	0.046
Stereotype Activation	22613	2	11307	4.869	0.009
Datatype * Stereotype Activation	1317	2	658	0.270	0.764

Table B.13: Results Post Hoc Test Datatype (Profile 6)

Comparison							
Datatype		Datatype	Mean Difference	SE	df	t	p_{tukey}
Sex-Swapped (Female)	-	Original (Male)	15.22	7.55	163	2.01	0.046

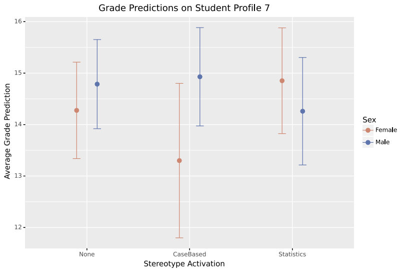
Table B.14: Results Post Hoc Test Stereotype Activation (Profile 6)

Comparison							
Stereotype Activation		Stereotype Activation	Mean Difference	SE	df	t	p_{tukey}
CaseBased	-	None	-28.0	8.99	163	-3.11	0.006
CaseBased	-	Statistics	-12.4	9.12	163	-1.36	0.365
None	-	Statistics	15.6	9.15	163	1.71	0.206

B.2.6.7 Profile 7

(Org. Boy with Grade = 13/20)
Reason School Choice - Close to home
Parents' education - High School
Studytime - between 2 and 5 hours
Absences - 0
Going out - Twice a week
Alcohol consumption - moderate
Freetime - low
In a relationship - yes

(a) Student Profile 7



(b) Average grade predictions

Figure B.11: Results on Student Profile 7

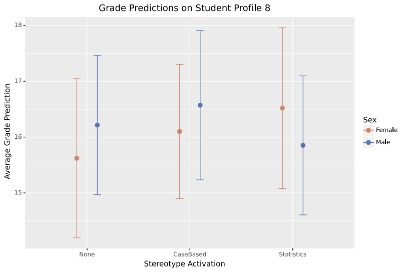
Table B.15: Results Statistical Test Profile 7

	Sum of Squares	df	Mean Square	F	p
Datatype	1626	1	1626	0.665	0.416
Stereotype Activation	281	2	141	0.057	0.944
Datatype * Stereotype Activation	7474	2	3737	1.557	0.214

B.2.6.8 Profile 8

(Org. Boy with Grade = 15/20)
Reason School Choice - Reputation
Parents' education - University
Studytime - more than 5 hours
Absences - 2
Going out - Once a week
Alcohol consumption - moderate
Freetime - high
In a relationship - no

(a) Student Profile 8



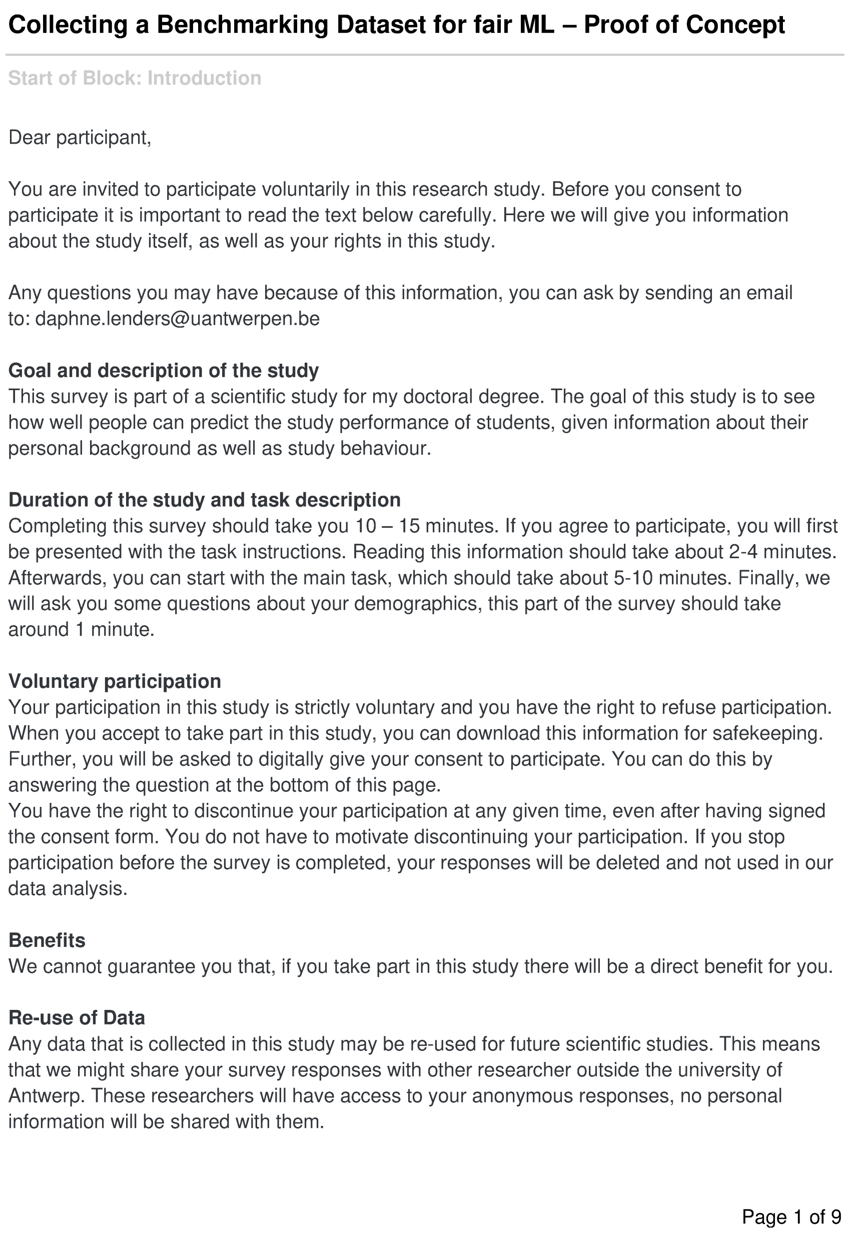
(b) Average grade predictions

Figure B.12: Results on Student Profile 8

Table B.16: Results Statistical Test Profile 8

	Sum of Squares	df	Mean Square	F	p
Datatype	49.5	1	49.5	0.0201	0.887
Stereotype Activation	1303.0	2	651.5	0.265	0.767
Datatype * Stereotype Activation	8253	2	4127	1.724	0.182

B.3 Survey Proof of Concept Study



Collecting a Benchmarking Dataset for fair ML – Proof of Concept

Start of Block: Introduction

Dear participant,

You are invited to participate voluntarily in this research study. Before you consent to participate it is important to read the text below carefully. Here we will give you information about the study itself, as well as your rights in this study.

Any questions you may have because of this information, you can ask by sending an email to: daphne.lenders@uantwerpen.be

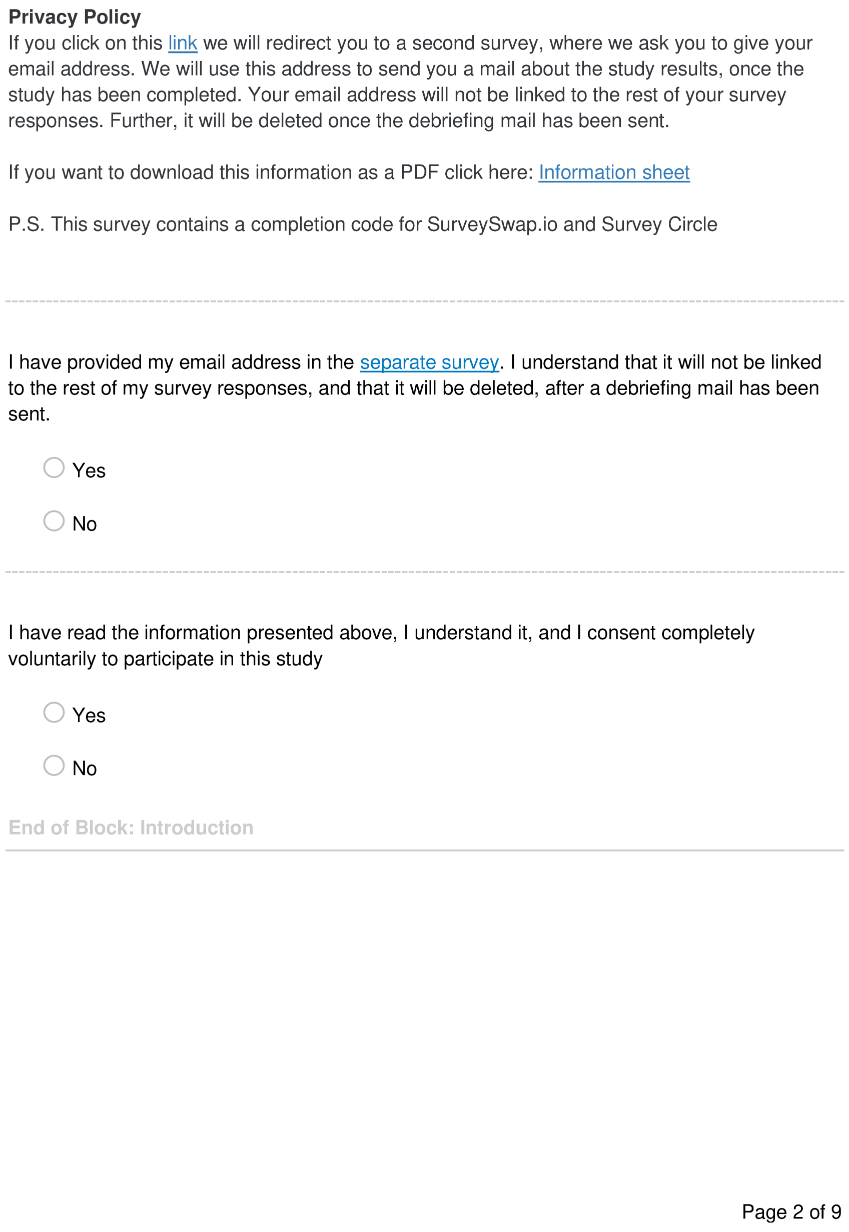
Goal and description of the study
This survey is part of a scientific study for my doctoral degree. The goal of this study is to see how well people can predict the study performance of students, given information about their personal background as well as study behaviour.

Duration of the study and task description
Completing this survey should take you 10 – 15 minutes. If you agree to participate, you will first be presented with the task instructions. Reading this information should take about 2-4 minutes. Afterwards, you can start with the main task, which should take about 5-10 minutes. Finally, we will ask you some questions about your demographics, this part of the survey should take around 1 minute.

Voluntary participation
Your participation in this study is strictly voluntary and you have the right to refuse participation. When you accept to take part in this study, you can download this information for safekeeping. Further, you will be asked to digitally give your consent to participate. You can do this by answering the question at the bottom of this page.
You have the right to discontinue your participation at any given time, even after having signed the consent form. You do not have to motivate discontinuing your participation. If you stop participation before the survey is completed, your responses will be deleted and not used in our data analysis.

Benefits
We cannot guarantee you that, if you take part in this study there will be a direct benefit for you.

Re-use of Data
Any data that is collected in this study may be re-used for future scientific studies. This means that we might share your survey responses with other researcher outside the university of Antwerp. These researchers will have access to your anonymous responses, no personal information will be shared with them.



Privacy Policy
If you click on this link we will redirect you to a second survey, where we ask you to give your email address. We will use this address to send you a mail about the study results, once the study has been completed. Your email address will not be linked to the rest of your survey responses. Further, it will be deleted once the debriefing mail has been sent.

If you want to download this information as a PDF click here: Information sheet

P.S. This survey contains a completion code for SurveySwap.io and Survey Circle

I have provided my email address in the separate survey. I understand that it will not be linked to the rest of my survey responses, and that it will be deleted, after a debriefing mail has been sent.

○ Yes

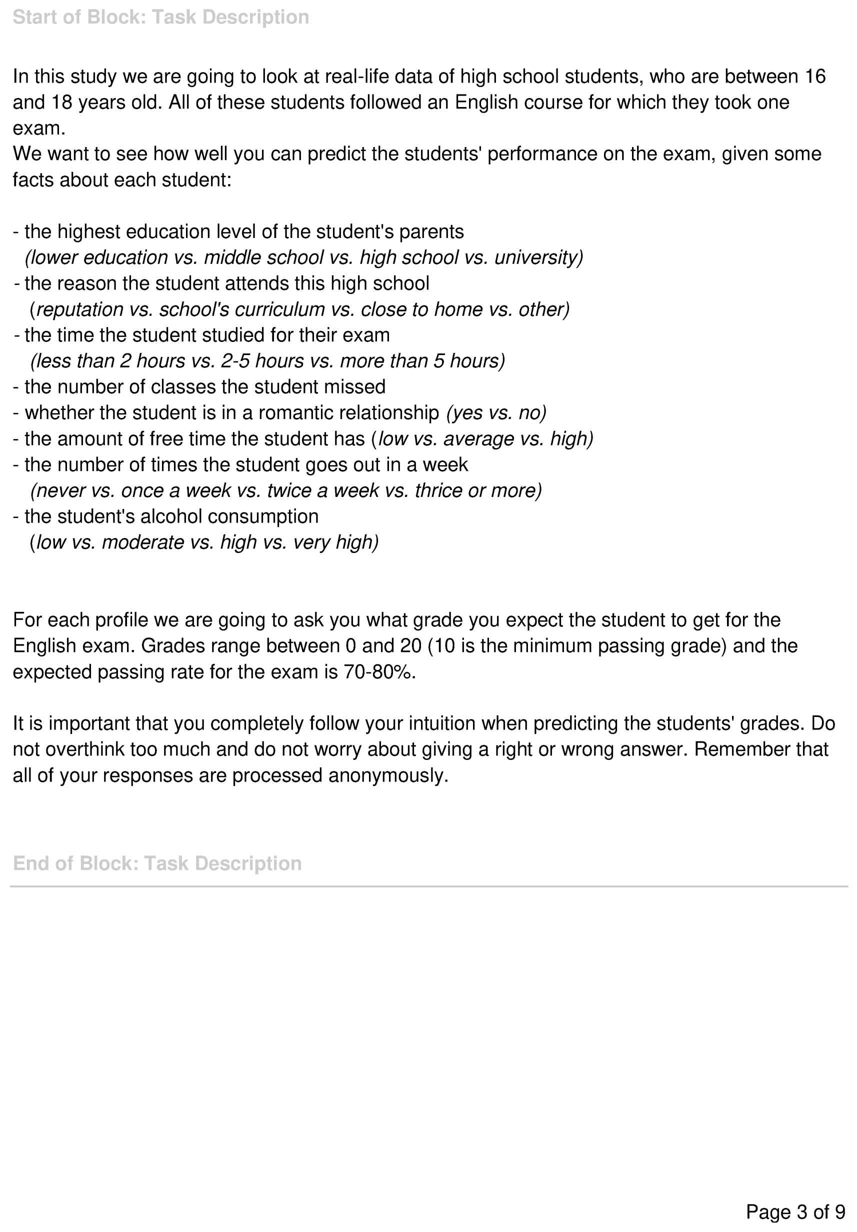
○ No

I have read the information presented above, I understand it, and I consent completely voluntarily to participate in this study

○ Yes

○ No

End of Block: Introduction



Start of Block: Task Description

In this study we are going to look at real-life data of high school students, who are between 16 and 18 years old. All of these students followed an English course for which they took one exam.
We want to see how well you can predict the students' performance on the exam, given some facts about each student:

- the highest education level of the student's parents
  (lower education vs. middle school vs. high school vs. university)
- the reason the student attends this high school
  (*reputation vs. school's curriculum vs. close to home vs. other)*
- the time the student studied for their exam
  (less than 2 hours vs. 2-5 hours vs. more than 5 hours)
- the number of classes the student missed
- whether the student is in a romantic relationship *(yes vs. no)*
- the amount of free time the student has (*low vs. average vs. high)*
- the number of times the student goes out in a week
  (never vs. once a week vs. twice a week vs. thrice or more)
- the student's alcohol consumption
  (*low vs. moderate vs. high vs. very high)*

For each profile we are going to ask you what grade you expect the student to get for the English exam. Grades range between 0 and 20 (10 is the minimum passing grade) and the expected passing rate for the exam is 70-80%.

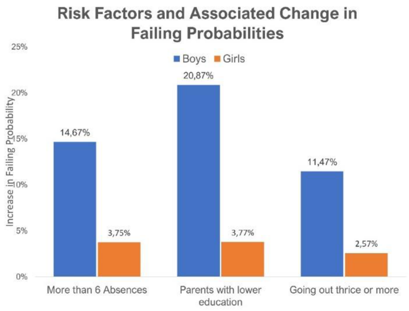
It is important that you completely follow your intuition when predicting the students' grades. Do not overthink too much and do not worry about giving a right or wrong answer. Remember that all of your responses are processed anonymously.

End of Block: Task Description

Start of Block: Statistics Stereotype Activation

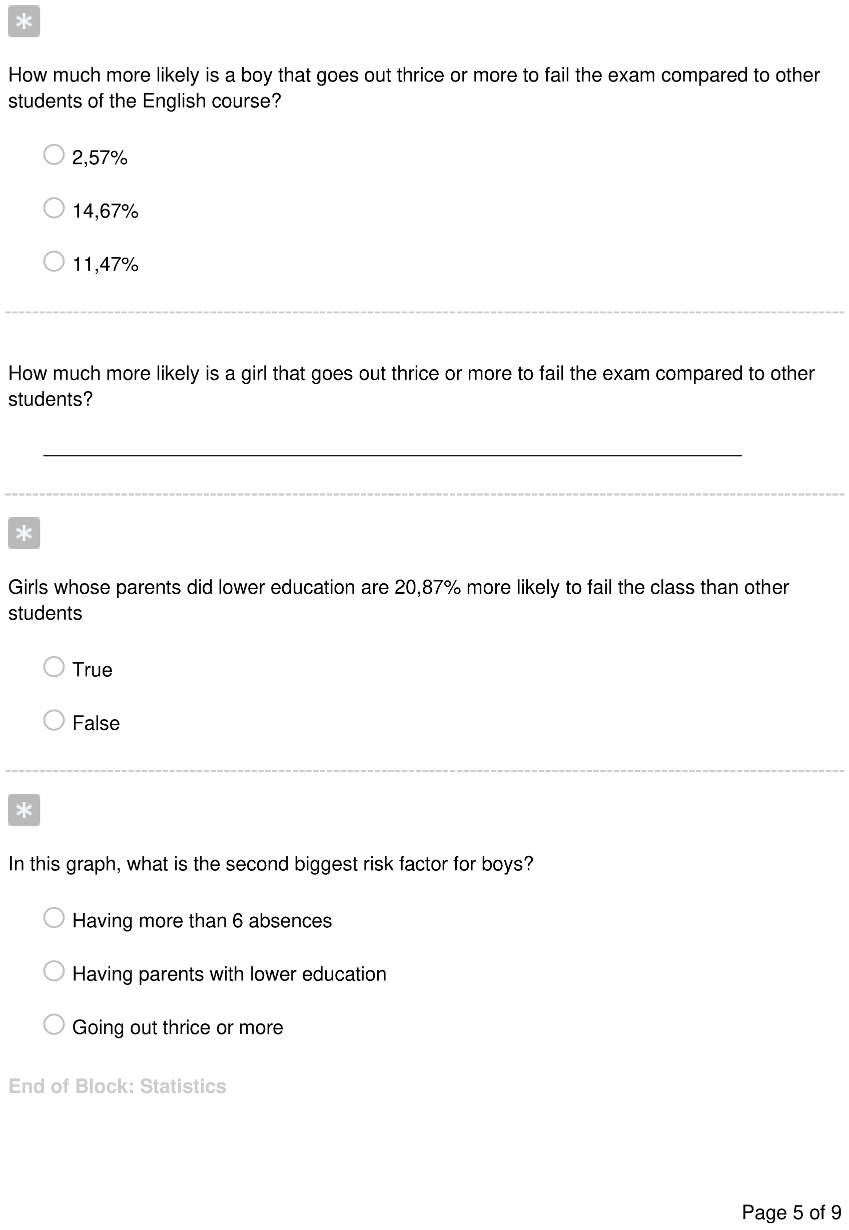
We already have some information about the associations between a student's characteristics and their chance to pass the English exam. More specifically, we identified some *risk factors*. Risk factors are characteristics of a student that are associated with a higher chance of failing the exam. One risk factor is, for instance, if a student goes out thrice or more per week. This student has a lower chance of passing the exam, than other students who go out less often.

Interestingly we found that risk factors affect boys more than girls. Take a look at the graph below where some risk factors are outlined:



To explain this graph look for instance at the risk factor "more than 6 absences": girls who fall into this category have a 3,75% higher chance of failing the course than other students. The effect for boys is bigger: if they miss more than 6 classes their chance of failing increases by 14,67%.

Please look at the graph for a bit and make sure you understand it. We are going to ask some questions to make sure you understand the effects of different risk factors.



*

How much more likely is a boy that goes out thrice or more to fail the exam compared to other students of the English course?

○ 2,57%

○ 14,67%

○ 11,47%

How much more likely is a girl that goes out thrice or more to fail the exam compared to other students?

____

*

Girls whose parents did lower education are 20,87% more likely to fail the class than other students

○ True

○ False

*

In this graph, what is the second biggest risk factor for boys?

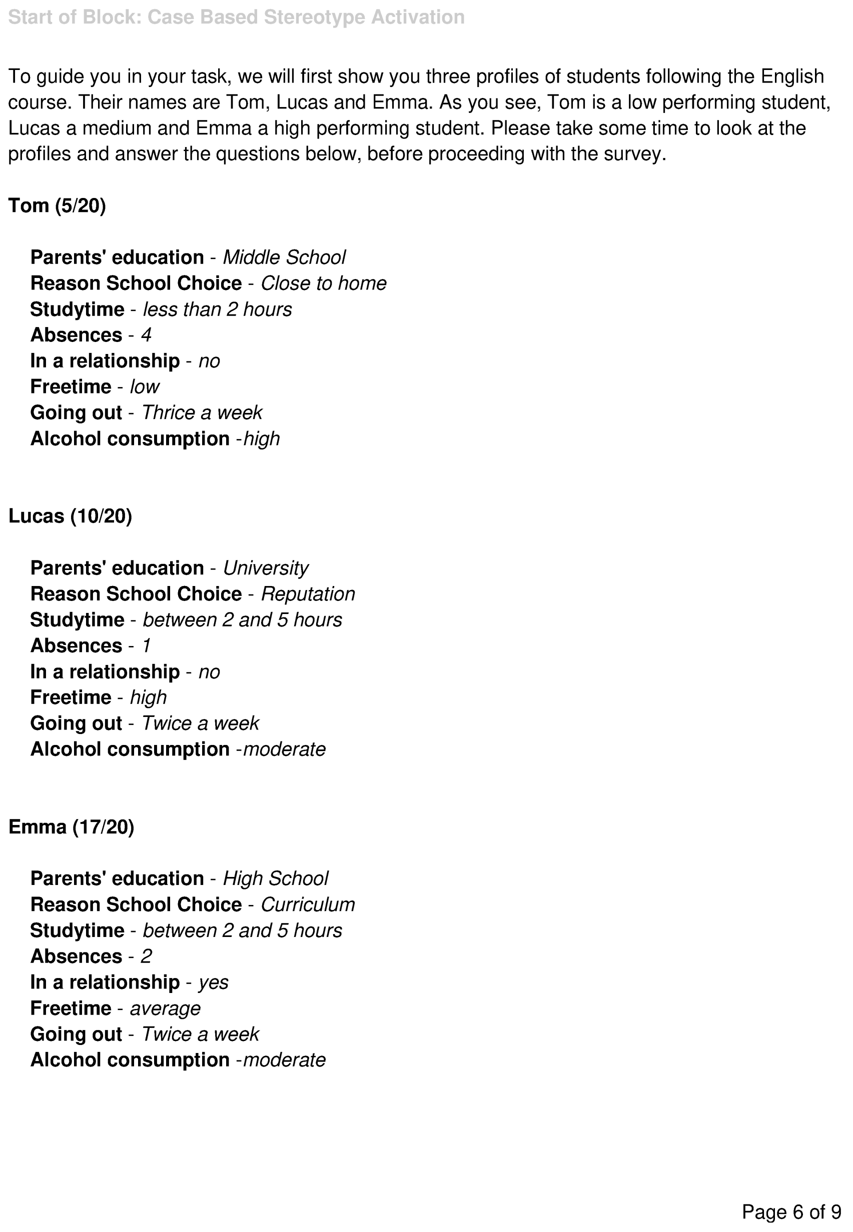
○ Having more than 6 absences

○ Having parents with lower education

○ Going out thrice or more

End of Block: Statistics

Page 5 of 9



Start of Block: Case Based Stereotype Activation

To guide you in your task, we will first show you three profiles of students following the English course. Their names are Tom, Lucas and Emma. As you see, Tom is a low performing student, Lucas a medium and Emma a high performing student. Please take some time to look at the profiles and answer the questions below, before proceeding with the survey.

Tom (5/20)

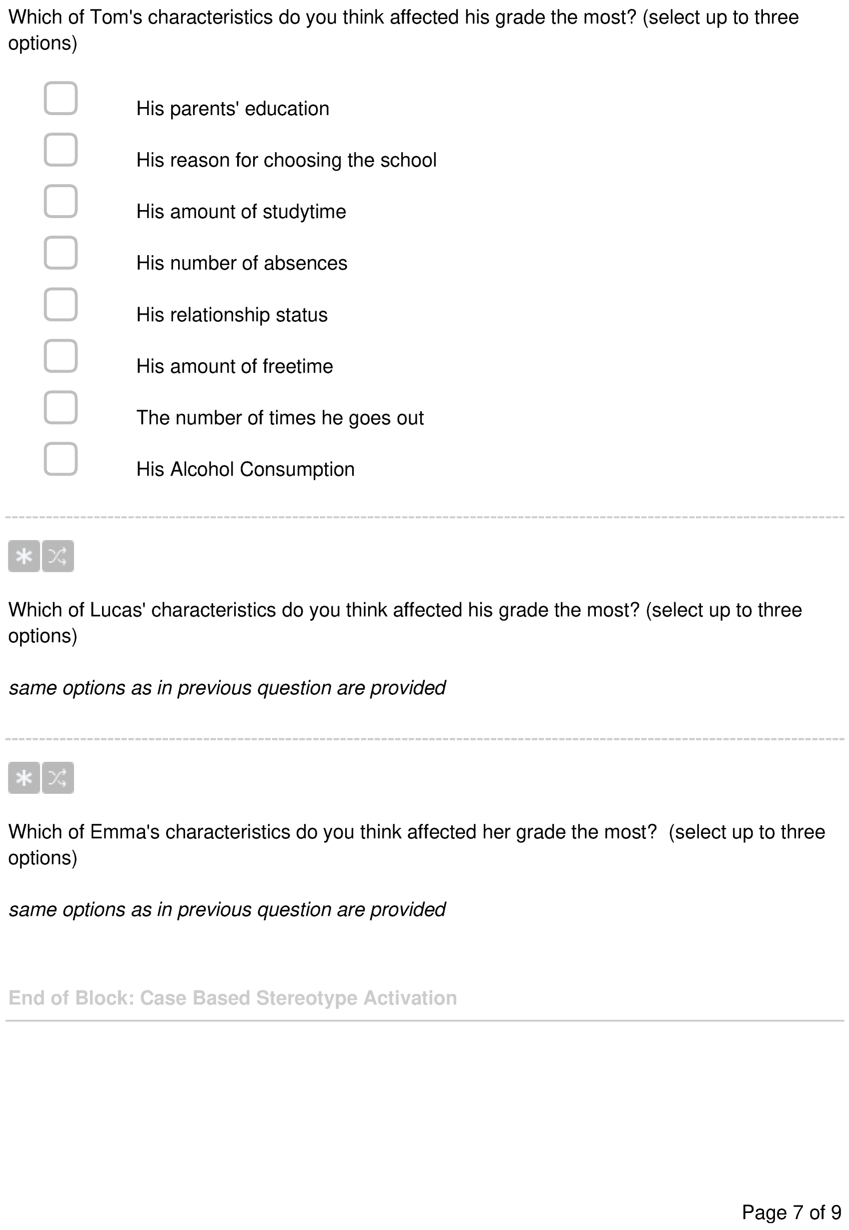
Parents' education - *Middle School*
Reason School Choice - *Close to home*
Studytime - *less than 2 hours*
Absences - *4*
In a relationship - *no*
Freetime - *low*
Going out - *Thrice a week*
Alcohol consumption -*high*

Lucas (10/20)

Parents' education - *University*
Reason School Choice - *Reputation*
Studytime - *between 2 and 5 hours*
Absences - *1*
In a relationship - *no*
Freetime - *high*
Going out - *Twice a week*
Alcohol consumption -*moderate*

Emma (17/20)

Parents' education - *High School*
Reason School Choice - *Curriculum*
Studytime - *between 2 and 5 hours*
Absences - *2*
In a relationship - *yes*
Freetime - *average*
Going out - *Twice a week*
Alcohol consumption -*moderate*



Which of Tom's characteristics do you think affected his grade the most? (select up to three options)

- □ His parents' education
- □ His reason for choosing the school
- □ His amount of studytime
- □ His number of absences
- □ His relationship status
- □ His amount of freetime
- □ The number of times he goes out
- □ His Alcohol Consumption

Which of Lucas' characteristics do you think affected his grade the most? (select up to three options)

same options as in previous question are provided

Which of Emma's characteristics do you think affected her grade the most? (select up to three options)

same options as in previous question are provided

End of Block: Case Based Stereotype Activation

Start of Block: Ranking Task

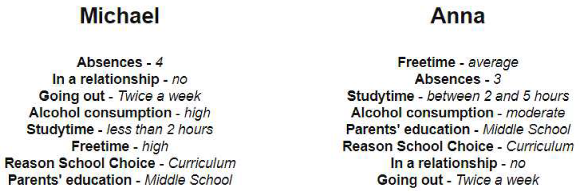
We arrived at the main part of this survey! Once you've completed this part, we will only ask you some more personal questions before you're completely done.

Below we present you with all student profiles. Please take some time to look at them and afterwards rank them according to how well you expect the students to perform in the English exam. Start with the student you think will get the highest grade.
You can use your mouse/touchscreen to drag and drop the student names.

For each student also specify in the blank field which grade you think they'll get. The grades range between 1 and 20 (10 being the minimum passing grade), and the expected passing rate for this exam is 70-80%.
Remember to completely follow your intuition and to not overthink your predictions.

_______ Anna
_______ Jenny
_______ Brian
_______ Oliver
_______ Sarah
_______ Michael
_______ Lisa
_______ David

Here all the student profiles are shown, to give two examples of a student profile



Michael

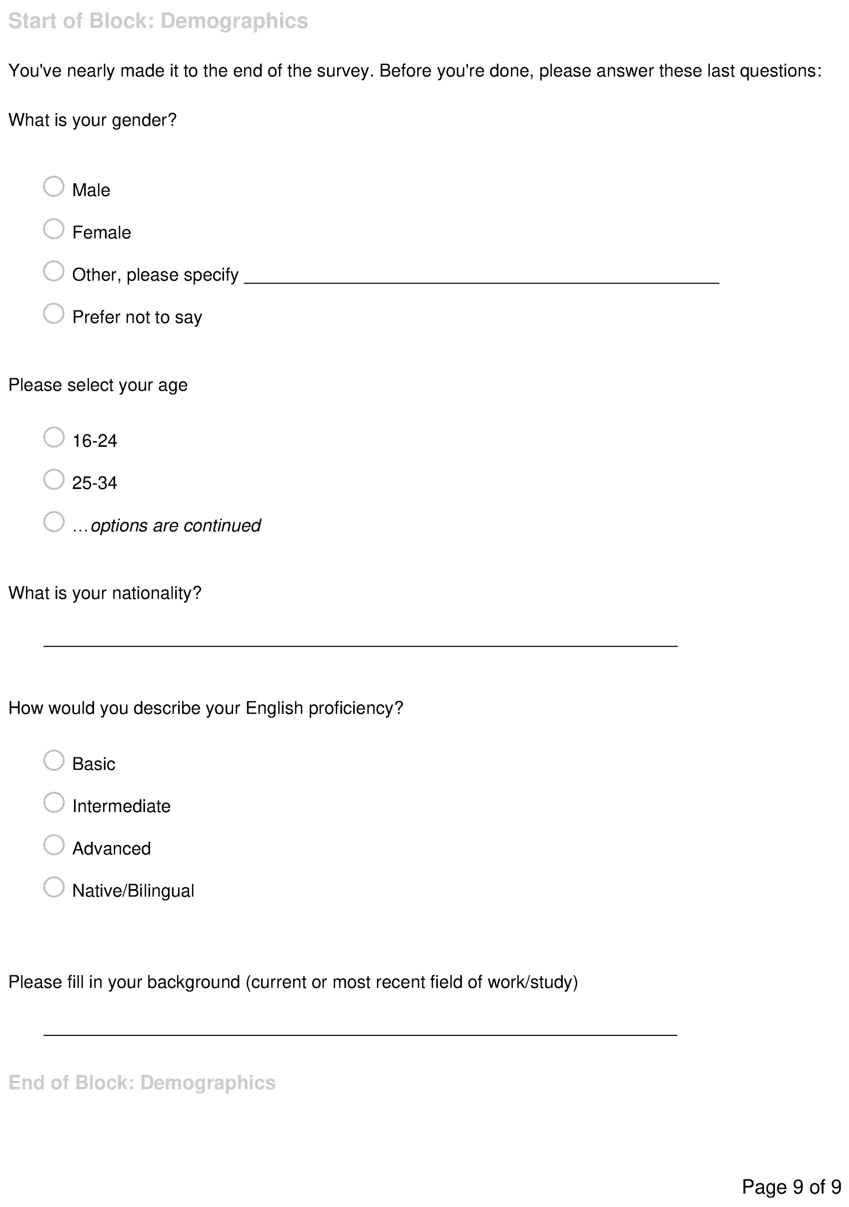
Absences - *4*
In a relationship - *no*
Going out - *Twice a week*
Alcohol consumption - *high*
Studytime - *less than 2 hours*
Freetime - *high*
Reason School Choice - *Curriculum*
Parents' education - *Middle School*

Anna

Freetime - *average*
Absences - *3*
Studytime - *between 2 and 5 hours*
Alcohol consumption - *moderate*
Parents' education - *Middle School*
Reason School Choice - *Curriculum*
In a relationship - *no*
Going out - *Twice a week*

(note in the other version of this survey the sex of the students will be swapped)

End of Block: Rank Table Format



Start of Block: Demographics

You've nearly made it to the end of the survey. Before you're done, please answer these last questions:

What is your gender?

○ Male

○ Female

○ Other, please specify ____

○ Prefer not to say

Please select your age

○ 16-24

○ 25-34

○ *…options are continued*

What is your nationality?

____

How would you describe your English proficiency?

○ Basic

○ Intermediate

○ Advanced

○ Native/Bilingual

Please fill in your background (current or most recent field of work/study)

____

End of Block: Demographics

Page 9 of 9

B.4 Survey Main Study

Because the survey we used for our Main Study was very close to that of our proof-of-concept study, we do not provide a copy of it (the survey used for the proof-of-concept study can be seen in Appendix C). As mentioned previously, the main change in this study lay in the presentation of the student profiles. Whereas in the proof-of-concept study each participant had to rank and grade the same eight student profiles, different profiles were presented per participant in our main study. Further, because the data from our main study is made publicly available, we added the following remark to our consent form:

> **Re-use of Data** The data that is collected in this study will be made available to public, so that it can be re-used for future scientific studies. This means, that researchers outside the university of Antwerp might access your anonymous responses. No personal information will be shared, however.

B.5 Results Main Study

B.5.1 Participants

In table we present the number of participants in our main study, distributed over the different "Stereotype Activation" conditions. In Table B.18 we show the number of participants by gender and age, while in Figure B.13 we show the number of participants by their nationality.

Table B.17: Number of Participants Across the Conditions

Stereotype Act.	Count	% of Total
None	32	30.2%
CaseBased	34	32.1%
Statistics	40	37.7%

Table B.18: Number of participants by gender and by age

Gender	Count	% of total
Female	74	69.2%
Male	32	29.9%
Prefer not to say	1	0.9%

Age	Count	% of total
18 - 24	68	63.3%
25 - 34	33	30.8%
35 - 44	5	4.7%
45 - 54	1	0.9%

B.5.2 Understanding the Effect of Stereotype Activation

In our paper we have shown how our collected benchmark data contains biases based on students' sex, that are mostly targeted against boys. However, this is only one of

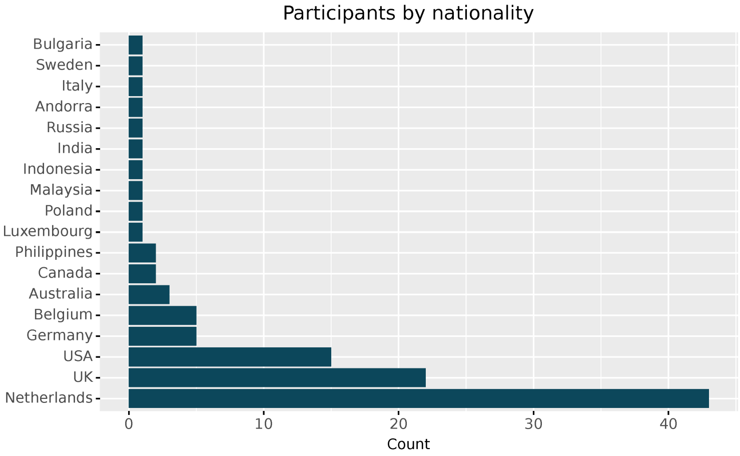


Figure B.13: Participants by Nationality

many introduced biases, and another interesting type of bias may stem from the type of stereotype activation that different participants were exposed to before the grading task. In particular, some of stereotype activation conditions may strengthen or weaken existing biases. To estimate these effects we used the binary version of the biased decision labels and compared how these relate to the actual labels per stereotype activation condition. In Figure B.14 we visualize the results.

One interesting observation from these plots is that the difference in discrimination rates between boys and girls is lowest if no stereotype activation is presented (difference of 5.31%), and highest for the “Statistics" condition (difference of 17.5%). The difference between these rates in the “CaseBased" condition lies with 11.76% between both. It is difficult to say whether these differences can completely be attributed to the various stereotype activation conditions, since participants in each condition got presented with different profiles and direct comparison for predictions on the same profiles is impossible. Still these numbers follow the patterns observed in the proof-of-concept study, and thus suggest that the different types of stereotype activation have some effect on how participants graded student profiles of different sexes. In regards to favouritism, it is hard to draw strong conclusions from the data. Relatively speaking, the difference in favouritism rates between boys and girls is highest when no stereotype activation is presented. However, given that not much favouritism occurs in the data overall, it is harder to observe clear patterns from this small fraction of the data.

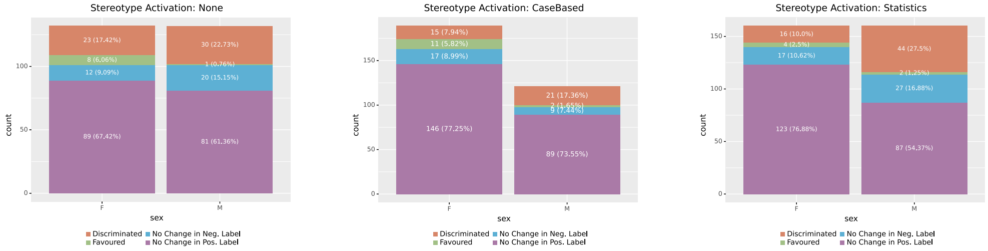


Figure B.14: Visualization of how different stereotype activation conditions affected the type/amount of bias introduced by our experiment.