Bias in, bias out: A Study on Social Bias in Automated Decision-Making Algorithms

A literature study on algorithmic bias, discrimination and fairness

Universiteit Utrecht

Author: Samaa Mohammad
First Supervisor: Michael De
Second Assessor: Benjamin Rin

BSc Artificial Intelligence
Faculty of Humanities
7,5 ECTS
17-07-2021

1. Abstract

Nowadays, algorithms play a large part in decision-making procedures, but they affect marginalized groups negatively when their decisions are driven by algorithmic social bias. An important way to look at this problem, is to investigate what notion of fairness marginalized groups need to be treated justly, and how to use this notion to find proper mitigation measures. This thesis aims to find how algorithmic bias in automated decision-making algorithms can be mitigated to prevent discriminatory decisions. In this context, algorithmic bias roughly refers to the concern that an algorithm is not merely a neutral transformer of data or extractor of information. There are many sources of algorithmic bias, and they emerge in different stages of machine learning.

The hypothesis was that researching the link between algorithmic bias, fairness and discrimination will help understand that mitigating bias is easily said than done, and that there will not be a one-size-fits-all solution for all cases. The results were found by performing a literature study on past research, and applying this knowledge to a pre-selected case which regards online automated proctoring algorithms. Exploring the proctoring case showed that it has no straight-forward way to mitigate its bias and achieving fairness requires a developer team to make trade-offs between different values.

The results suggest that there is no coherent answer to how to mitigate algorithmic bias, and that mitigation measures are case-specific. However, it seems that collecting representative datasets, developing an algorithm in a diverse team and creating it *with* misunderstood groups could generally help mitigate algorithmic bias.

Table Of Contents

2. Introduction

Algorithms are quickly becoming the main instrument for decision making procedures. Many aspects of society are affected by so called automated decision making algorithms: data-driven algorithms that output automated decisions without human involvement. Due to their accuracy and efficiency, automated decision making algorithms are widely applied to fields such as hiring [1]–[3], education [4], [5] and criminal justice [6], [7], operating at a large scale. Although it is expected for algorithms to be objective and fair, it has been observed that they are biased and place certain groups of people at a systematic disadvantage. These are cases of discrimination – which is legally defined as the unfair or unequal treatment of an individual (or group) based on certain characteristics such as religion, nationality, familial status, age, sex, income, education, gender and ethnicity. [8] One of the greatest challenges within Artificial Intelligence (AI) is combatting digital discrimination. A well-known example of digital discrimination is a consequence of the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm that was used in US court systems wherein the algorithm predicted twice as many false positives (misclassifications) for recidivism for black offenders than white offenders [9]. Another infamous case involves Google Photos, which tagged two African-Americans as gorilla's through their facial recognition software [10]. While not directly linked to automated decision-making, the appearance that facial recognition software often fail to identify people of colour could have drastic effects when an automated decision-making algorithm relies on such a mechanism.

It is now well established from a variety of studies that bias is inevitable, but the majority of literature focusses on researching proper mitigation measures per case. Several researchers have been exploring the intersection between bias, discrimination and fairness. Still, much uncertainty exists about this, as researchers propose different notions of fairness to combat algorithmic bias.

The purpose of this thesis is to review recent research into the field of algorithmic bias and discrimination. Therefore, this thesis aims to answer the following research question: **'How can algorithmic bias in automated decision-making algorithms be mitigated to prevent discriminatory decisions?**' To properly answer this research question, in this thesis we will go over notions of algorithmic bias, fairness and social discrimination that are applicable within a pre-selected, relevant and actual case from practice, namely the use of online automated proctoring algorithms. The case is selected based on personal and timely relevance in the current COVID-19 pandemic, and the recent backlash it received. The hypothesis is that researching the link between algorithmic bias, fairness and discrimination on this will help understand that mitigating bias is easily said than done, and that there will not be a one-size-fits-all solution for all cases.

The thesis has been organised in the following way. This thesis begins by describing algorithmic bias (chapter 2). It will then go on to evaluating the problematic side of bias. The remaining part of this thesis proceeds in the following way: it will begin by evaluating one case, describing the source of algorithmic bias, fairness, discrimination, and possible mitigation measures, and a brief review of what we need besides a fair algorithm to prevent discriminating algorithms (chapter 3). In chapter 5 this thesis will be concluded.

3. On algorithmic bias

This chapter will discuss the definition of algorithmic bias and introduce different kinds and sources. Per type of bias, it will show some examples and discuss how these biases can become problematic. Moreover, it will include a discussion on how such bias can lead to discrimination.

3.1. What is algorithmic bias?

The most general term of bias refers to systematically making decisions or judgements based on prejudices. Long before the current debate about algorithmic bias in machine learning, Friedman and Nissenbaum [11] pioneered in the researching field of bias in computer systems and stated that a computer system is biased if they "*systematically* and *unfairly discriminate* against certain individuals or groups of individuals in favor of others [by denying] an opportunity for a good or [assigning] an undesirable outcome to an individual or groups of individuals on grounds that are unreasonable or inappropriate". Here the definition of bias focusses on its moral implications. For example, a school can be 'biased' towards children from minorities, which leads them to exclude the children from educational programs.

However, bias does not imply discrimination at all times. Another widely cited definition of bias in computer science is formulated by Danks and London [12] in a more neutral way: "Bias means a deviation from the standard, sometimes necessary to identify the existence of some statistical patterns in the data or language used." [8] For example, statistical bias refers to an estimate that deviates from a statistical standard and moral bias refers to a judgement that deviates from a moral norm. Social bias, legal bias, and many other types of bias are likewise a deviation from their type of standard [12]. It is important to note that bias itself is not something to be immediately discarded. It plays an important role in classifying and finding differences between instances [8]. Also, it is important to note that there can be bias in one respect (e.g. moral), but not another (e.g. statistical). For example, women are statistically less represented in professions dominated by men (e.g. such as construction working, data engineering and mechanical engineering). This shows a deviation of the statistical standard (the true population value), which is characterized as statistical bias. According to Danks and London, such statistical biases often serve as a tool to identify moral biases. They argue that the underrepresentation of women in male-dominated professions may raise questions about unobserved, morally problematic hardships women have who work in these professions [8] .

From this definition it can be understood that bias is inevitable. Bias is used to continually make judgements and decisions and does not always lead to discrimination. However, social bias towards protected attributes (e.g. gender, race, age, religion, etc.) is questionable as it is hard to say what the objective 'standard' is, and who deviates from it. For example, people affected by gender bias usually receive different treatment based on

the person's real or perceived gender identity [13]. This thesis will mainly focus on exploring social bias in algorithms.

Danks and London argue that there is no coherent notion of 'algorithmic bias' in most cases. Many cases contain multiple sources of bias, each with their own nature, that require different responses. For example, the COMPAS case, the algorithm that predicts recidivism, is affected by algorithmic bias in multiple ways: through legal, moral, racial, and statistical bias. Each of these biases require its own proper mitigation measures and analysis [9], [12]. Moreover, in the next section it will be clear that the biases can be introduced through different stages of a machine learning project.

Although uncertainty of the precise definition still exist, there appears to be some agreement that algorithmic bias refers to "roughly, the worry that an algorithm is, in some sense, not merely a neutral transformer of data or extractor of information" [12]. This thesis is supported by this sentiment in combination with the described notions of social bias and social discrimination.

3.2. What kinds of sources for algorithmic bias exist?

Multiple researchers have provided taxonomies to distinguish between different sources of algorithmic bias. This thesis explores two suggested taxonomies by widely cited researchers Danks and London [12] and Barocas and Selbst [14], which form a basis for most research on algorithmic bias. With these taxonomies in mind, this thesis aims to find the relevant sources of bias wherein social bias in automated decision-making can infiltrate or occur and possibly cause discrimination. Therefore, this thesis focusses on a selection of the sources of bias in the taxonomies. These types and sources will be highlighted: defining the 'target variable' and 'class labels', training data, feature selection, proxies [14], algorithmic focus bias, algorithmic processing bias, and transfer context bias [12]. It is not the task of this thesis to examine every existing source of bias, nor did the researchers intend to provide an exhaustive list. Thus, this thesis chooses to show an overview of biases that possibly lead to discriminatory decisions. The following subsections attempt to make clear where the bias comes from, how problematic bias can be identified, and how it leads to problematic decisions. The biases will be distinguished between bias in modelling, training and usage.

3.2.1. Bias in modelling

Defining the 'target variable' and 'class labels'

Programming a machine learning algorithm takes multiple steps, where the first step is to model the data to match the goal one has in mind. In the modelling phase, the relevant variables are selected, data is curated and

prepared into a comprehensive set of data that can be used to train an algorithm to make decisions and predictions with new data. In the modelling phase, Barocas and Selbst identified a source of bias that comes into existence through wrong definitions of *target variables* and *class labels* [14]. The target variable in this phase here refers to the desired output of the beneficiary of the algorithm; a speaking example for hiring algorithms can be the definition of a potentially 'good' employee. This target variable has not an obvious definition, but it is up to the developer team together with the employer to specify what a 'good' employee characterizes. The definition of 'good' is specified in the class labels, where all possible properties the target variable are allocated to categories [14].

The specification of the target variable into class labels is a subjective step. Although one might see that the specification of a 'good' employee could be defined with objective, measurable properties, such as sales rates and work efficiency, these measurable properties cover only a subset of the broad definition of 'good' [14], [15]. Thus, an employer might prefer to add subjective class labels, e.g. not coming late to work and being good at teamwork. Barocas and Selbst argue, however, that "while different choices for the target variable and class labels can seem more or less reasonable, valid concerns with discrimination enter at this stage because the different choices may have a greater or lesser adverse impact on protected classes." [14] For instance, the class label of not coming late to work can reflect bias. Consider a situation wherein people with low income rarely live close to the city center. They would have a longer commute to work than people with more income, who might live closer to the city center. Commuting longer gives a higher chance of encountering obstacles in transport (e.g. traffic jams, train delays, etc.). As people with low incomes have bigger chances of coming late to work than people who live in the city center, they would not be attractive to an automated hiring algorithm used by a company that is located in the city center. Defining a 'good' employee with the class label of not coming late to work can do harm to those of protected classes, as people with low income would be unintentionally discriminated against [15].

Feature selection

To reduce the computational costs and sometimes also the performance of a predictive model, it can be beneficial to reduce the amount of variables that are used to train and operate the model [14]. This is done in the process of *feature selection*. A feature[1] is "an individual measurable property or characteristic of a phenomenon" [16]. Barocas and Selbst argue that using wrong or too superficial features may lead to a biased model, causing harm towards people of protected groups [14]. They explain the concern that if the attributes that make the variation within the protected group clear are not included, the model may not be able to distinguish group members, causing it to rely on broad generalizations that are detrimental to individuals of the protected class [14], [17].

[1] We use the terms variable, attribute and feature here interchangeably.

In a simplified example, employers might assign enormous weight to the reputation of the university from which a candidate has graduated from, because they associate highly accredited universities with excellent graduates and thus good employees. The feature here is the reputation of an university, which is cheap and easy data to obtain as it is available in the candidate's resume [14]. However, the reputation of an university does not guarantee that its graduates are competent and excel in employability skills [2]. Nevertheless employers continue to hire graduates from excellent universities, which usually have a high tuition fee. However, due to their high tuition fees, excellent universities rule out students who have low socioeconomic status. Often, this group consists of immigrants and ethnic minorities. Thus, the employers choice of feature for the automated hiring algorithm leads to indirect discrimination to those with little financial backing, particularly ethnic minorities [18].

Even when students of protected classes happen to graduate from such universities, they would only form a minority. So when they are equally competent to their non-minority peers, the hiring algorithm will, according to Barocas and Selbst, still "incorrectly and systematically discount" [14] minority individuals, if it bases its decision on the accreditation feature instead of the distinct, more accurate and holistic qualities of the candidates [14]. In other words, if the attributes that explain their individual distinction are not included in the set of features, the model will draw its decision on broad generalizations, which is harmful towards protected groups as generalizations and stereotypes are often rooted in other sources of algorithmic biases.

An algorithm seldomly weights one feature so heavily that it alone will be determining for the outcome of it, but this is a way in which certain protected groups will be disadvantaged over others.

Algorithmic focus bias

Algorithmic focus bias occurs when some data attributes get more emphasis in the model than relevant [12]. It can be that the model focuses on race, while race is irrelevant to making a good decision.

One recent infamous case wherein algorithmic focus bias led to discriminatory decisions, it that of the Dutch childcare benefits scandal (Dutch: toeslagenaffaire). The toeslagenaffaire has received much public attention since 2018, when investigators discovered that the Dutch Tax and Customs Administration had been falsely alleged 26,000 parents of fraud with the aid of a risk assessment algorithm between 2013 and 2019 [19], [20]. The governmental organization claimed the parents' received benefit claims back in their entirety, which caused many of them to become bankrupt as the debt amounted tens to thousands of dollars [21]. This situation showed that the Tax and Customs Administration made severe mistakes, while being accused of discriminatory decisions. Together with conspiracies within the cabinet and evidence of institutional bias, the cabinet decided to resign over the scandal, because "fundamental principles of the rule of law" had been violated [20]–[22].

One of the reasons why the Dutch cabinet concluded that the toeslagenaffaire was a case of discrimination, is because the Dutch Data Protection Authority (AP) published a technical report [23] on this automated decision-making algorithm in 2020. They explain that the algorithm uses a risk assessment model, which predicts the likelihood someone will commit fraud. The main point of their research is that the algorithm unnecessarily processed dual citizenship in their risk assessment model, which is a protected attribute. They concluded that the algorithm focused too negatively on this attribute, whereas it was not necessary to include for the algorithm to fulfill its task [23]. The chairman of the AP points out: "The whole system was set up in a discriminatory way and was used as such. [...] There was permanent and structural unnecessary negative attention for the nationality and dual citizenship of the applicants." [24] Thus, this example of algorithmic focus bias shows that when a protected attribute gets emphasized on more than necessary, discriminatory decisions can occur.

Proxies

Programmers choose to include relevant variables at the creation of rational and knowledgeable automated decision-making algorithms. Programmers who intend to protect privacy and avoid discriminatory decisions may choose to exclude variables that categorize people on their protected attributes [25]. However well intended, algorithms can still make decisions that disadvantage people of marginalized groups by using *proxy variables*. Arbitrary variables can match protected variables when they are highly correlated to each other. In other words, algorithms can still make biased decisions when proxies serve for protected variables [14], [25]. For example, zip codes are commonly included in datasets, while it serves as a proxy for race [26], [27]. Reason for this is that zip codes are historically strongly correlated to the racial composition of residential areas [28]. Thus, zip codes can, despite seeming a neutral variable, cause racial discrimination. Other neutral variables that can cause discrimination can be as simple as the purchase of certain products that are highly popular among groups of people. For instance, the consistent purchase of halal food is a proxy to religious beliefs, as Muslims are obligated to consume halal food only. A supermarket could use this proxy to adjust their pricing or marketing strategy [27]. Zarsky [27] warns that such a proxy for religion could be simply inserted into a scoring mechanism, which will reflect biases and discriminate against people with certain religious beliefs. He argues that "the use of such factors is socially unacceptable, and that these factors must be removed from the scoring algorithm when highly correlated with a protected group." [27] However, there seems to be no consensus on how to treat proxies. In his book 'Understand, Manage and Prevent Algorithmic Bias', Baer [29] argues that if you remove a proxy, there will be another proxy to substitute for that proxy. Williams, Brooks, and Shmargad [25] plead for not removing protected attributes at all to prevent proxies from impacting algorithmic decisions. In their article, they show that discrimination is harder to ascertain when it is caused by proxies and bias

is more difficult to detect, whereas protected variables can be proactively used to help mitigate and combat discriminatory decisions. [25] The next chapter will suggest some different mitigation measures for algorithmic bias along with their suggestions.

All in all, it is important to keep in mind that the exclusion of protected variables may nevertheless exacerbate, instead of prevent, discrimination.

3.2.2. Bias in training

Training data bias

An algorithm is built with the use of a dataset. The dataset is often split in training data and in testing data. The prior is to train the model and the latter to validate whether the input variables and the algorithm 'learned' to come to the correct outputs. The algorithm requires a complete dataset with accurate and objective labeled data [30]. However, within the process of data labelling and collection, multiple biases can be found.

First, if a dataset contains data of past decisions that are motivated by prejudices and social bias, and those are labeled as valid examples, the algorithm will learn to adopt the same behavior [14]. Veale & Binns [31] point out that "if these historical data reflect existing, unwanted discrimination in society, the model that is learned from it – essentially a similarity engine – will likely encode these same patterns, risking reproduction of past disparities." For instance, audit studies – a type of study used to test discriminatory behavior – have shown that women and ethnic minorities are often discriminated against in the job application process [1], [3], [32]. To show that existing prejudice and bias in an already existing system can reproduce bias and discrimination, consider the following case. From 2014 until 2018 Amazon delegated its resume reviewing process to an automated recruitment algorithm. Applicants were ranked through a rating system (giving applicants a score from 1 to 5 stars), and the applicants with the best score would be hired to fill software developer positions. However, the algorithm's training dataset was dominated by resumes of men over a ten-year period. The algorithm downgraded resumes that mentioned women's colleges, and "penalized resumes that included the word "women's," as in "women's chess club captain."" [33], [34]. It is important to note that the tech field has been dominated by men. So, not only were women underrepresented in the dataset – which made them less likely to be hired by the algorithm –, but also perceived as unqualified based on their gender. The algorithm reproduced prejudice of prior decision makers, reflecting bias that already exists in the hiring process [14]. This led to the algorithm labeling those decisions as valid examples.

According to Danks & London, it is not easy to identify this type of social algorithmic bias. The bias can slip unnoticed into the training data, and remain subtle and hidden. Most of the times, the algorithms are opaque

as developers keep the precise training dataset undisclosed. They state: "If we only see the final learned model or its behavior, then we might not even be aware, while using the algorithm for its intended purpose, that biased data were used." [12]

Second, if a dataset contains a biased sample of the population, any decision that affect those who are under- or overrepresented in the dataset may put them on a systematic disadvantage [14]. For example, in supervised learning, a developer team may assemble a dataset of faces where all faces are labeled as such. However, when one is unaware of the homogeneity of the dataset (i.e. only containing white faces), it may occur (as experienced by professor Sennay Ghebreab [35]) that the algorithm used to decide to open an automated door will refrain black faces from entering the building. Another well-known example involves the facial recognition software of Google Photos, which tagged African Americans as gorillas [10]. In these two cases, underrepresentation of people with dark skin led to discrimination and exclusion towards them.

On the other hand, overrepresentation has its own negative effects. Consider a predictive policing algorithm – an algorithm used to identify potential criminals. Such algorithms rely on historic data of past criminal activities, attempting to predict a criminal profile. Minorities that are already institutionally discriminated against receive unjustified convictions more often than the dominating population [36]. The combination of this and the idea that the police arrests minorities more often, leads negative information regarding minorities to be overrepresented in the historical data. In a similar case for minority hiring, Zarsky [27] concludes: "This might result from an oversampling bias; given existing prejudice, minorities are sampled more often for indiscretions, and thus their indiscretions are over-represented in the database." [37], [38] These are concrete examples of poor data collection that led to non-representative training datasets. It is possible, therefore, that allowing training data bias to hatch can lead to the creation of discriminatory models [1], [39], [12], [14], [40].

3.2.3. Bias in usage

Transfer context bias

Transfer context bias occurs when an algorithm is applied outside of its intended context of use [12]. This bias is more a user bias rather than a bias inherited by the algorithm itself. Digital discrimination can still occur, but it is not caused by the algorithm. Consider an hypothetical algorithm[2] intended for recognizing and recruiting male athletes for the Olympics. The algorithm is specifically designed for male athletes, so it is looking at the distinct qualities for a male to be qualified for the Olympics. Female athletes, however, have different requirements in order to be qualified, and

[2] A real life example in the context of automated decision making was difficult to find.

compete separately from men. So when the previously described algorithm is applied to women, it is very possible that the algorithm will be biased towards them, because they have distinct bodily features and qualities when compared with men. They will be left out of the competition as they don't meet the criteria of the algorithm. This example of transfer context bias shows that an algorithm that is designed for a specific context of use should not be used inappropriately in a different context that it is not intended for. Moreover, the described case is not an example of a discriminatory decision by an algorithm, rather than inappropriate use of an algorithm. Most of the cases wherein processing bias leads to problematic automated decisions is due to user error [12].

4. Fair algorithms

This chapter will introduce the pre-selected case that will help answer the research question. The first part will describe a brief introduction to fairness. Then it will proceed to discuss the online automated proctoring case, describing its sources of bias, notion of fairness and potential mitigation measures. This chapter will end with a short overview of additional general mitigation measures.

4.1. Brief introduction to fairness

Throughout this thesis, one might question: what makes an algorithm *fair*? 'Fairness' is a term frequently used in machine learning literature, but to date there is no consensus about its notion. In fact, the debate around its notion has a long history in philosophy, long before machine learning literature on fairness started rising [41]. The term could be broadly described, however it is out of the scope of this thesis to explore the full philosophical definition of fairness. Instead, the exploration of fairness will focus on finding a notion to support mitigation measures against discriminatory decisions for the selected case.

With respect to the definition of fairness in automated decisions-making, there appears to be some agreement that fairness refers to "the absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics." [41] Friedman and Nissenbaum's definition of bias (see chapter 1) align with this agreement. In other words, it is agreed that fairness refers to unbiased treatment.

However, there exist many ways to achieve fairness. In machine learning, many frameworks and mathematical formulations have been proposed to accomplish fair machine learning algorithms (see e.g. [42]–[48]), but these frameworks need to be interpreted with caution. Every case needs to be treated differently, as there is not a one-size-fits-all solution to each of them, because fairness needs to be defined for specific context and applications. It is important to first determine whether fairness should be received on an individual or group level [41]. Moreover, according to Kleinberg, Mullainathan and Raghavan [49] it is impossible to satisfy different fairness notions at once.

To illustrate the diversity of fairness notions, a few examples follow: fairness could be achieved through awareness, which means that "an algorithm is fair if it gives similar predictions to similar individuals" [41], [42]; equal opportunity, which implies that "the protected and unprotected groups should have equal true positive rates" [41]; treatment equality, which is achieved "when the ratio of false negatives and false positives is the same for both protected group categories" [50]; counterfactual fairness, which is the "intuition that a decision is fair towards an individual if it is the same in both the actual world and a counterfactual world where the individual belonged to a different demographic group" [41], [51]; accuracy equity, which considers the overall accuracy of a predictive model for each group [31], [52]; or equalized odds, which states that "the protected and unprotected groups should have equal rates for true positives and false

positives" [41]. Others include fairness through unawareness, conditional statistical parity, and test fairness (see [31], [41]).

Specifying the notion of fairness is an important step to understand what bias mitigation techniques should be used to prevent discriminatory decisions. The question one should ask throughout examining every case is, *why* does this algorithm make unfair decisions, *what exactly* is unfair about this discriminatory result, and *how* can a specified notion of fairness be used to select mitigation measures? The following section attempts to examine automated online proctoring algorithms, while focusing on these questions by describing its bias, fairness and mitigation measures. However, the results will be hypothetical, as not much research has been done on this very recent topic. This thesis will use reasoning and findings to unravel its problematic decisions.

4.2. Case: online proctoring algorithms[3]

4.2.1. Introduction and problem statement

Due to the effects of the COVID-19 pandemic, higher education (among other work fields) has shifted towards operating almost fully online. Consequentially, students are attending online classes and taking exams via online software. The rise of online exams has brought concerns on the difficulty of monitoring students during exams, leaving teachers to blindly trust on the honestly of their students. However, students always find ways to cheat. For example, students are able to cheat by discussing questions and answers on group chats, letting someone else take the exam on their behalf, taping notes on their screen, and switching to another monitor to look up information [53].

In respond to the ease of online cheating, universities have experimented using proctoring services, such as Proctorio [54], Respondus [55], Honorlock [56] and ProctorU [57], which track students in real time through their webcams and laptop screens. These services offer both live individual human-proctoring, and automated proctoring. For both methods, students are required to identify themselves using their ID, make a video of their surroundings, leave their microphone unmuted, and be recorded. During live human-proctoring, an employee of the proctoring service monitors a student individually and flags cheating behavior, leaving it to be judged by the course administrator. On the other hand, automated proctoring uses algorithms to detect cheating behavior in real time. Again, the course administrator makes the final judgement. Proctoring algorithms use techniques as AI (to detect voice using voice recognition to recognize dishonesty), machine learning (to train on correct flags) and biometrics (including facial recognition to identify students with; eye tracking to track whether a student gazes on their laptop screen; and facial detection to determine the presence of the student) [58].

[3] After establishing the results independently, I discovered a paper by Coghlan, Miller and Paterson [79], wherein their findings converge with mine.

However, these proctoring algorithms did not come without problems. Several students have reported that they were falsely accused of cheating and consequently had their exam rejected for reading a question out loud [59], having concentration difficulties, and having high rates of eye movement due to their neurodiverse nature [60]. Moreover, a major group of black and brown students were asked by the service to shine a bright lamp on their face, despite being in a well-lit room, because the facial recognition software could not recognize their face in the identification stage. During their entire exam, the bright lamp caused headaches for several students [58], [61]–[63]. This facial recognition software caused trouble for (closeted) transgender students as well, as their new appearance, gender and name often do not correspond to the information on their ID, leaving them in a discomforting situation before the start of the exam [60].

An additional concern about online proctoring algorithms regards privacy exceedance. First, many proctoring software require students to film their entire desk and room before starting an exam [54]–[57]. After that, their room will still be visible in the background, which many students experience as undesirable [64], [65]. Second, some require students to remain unmuted [56]. This raises privacy and autonomy concerns, as students are constantly being monitored in their private room while being restricted to leave the seat. Moreover, conversations in the background can be picked up by the microphone, which will be found back in the recording. It is out of the scope of this thesis to discuss further privacy concerns as it does not add to the research question, but it is worth mentioning the problem.

Critical questions around the set-up of the algorithm arise. When should it accuse a student of cheating? Do companies behind these algorithms take into account the diversity of students? What about the historical flags in the training dataset that were grounded on unjustified, human error punishments? What could be ableist, unfair and discriminatory about online proctoring algorithms?

4.2.1.1. Type of bias

Chapter 1 described different types of algorithmic bias. Automated proctoring has many different problematic issues, but for the sake of addressing the research question of this thesis, the focus will lie on the problems that cause biased decision-making. The most prominent bias is towards ableism (not taking neurodiverse – that is diagnosed and undiagnosed with e.g. dyslexia, autism, ADHD, and other learning disabilities - and disabled people into consideration) [66].

It can be difficult to find the true source of ableist bias in proctoring algorithms, although this thesis will make an attempt to reason towards a source. While the tasks of its eye tracking mechanism is to track whether students gaze appropriately to their exam, it is merely doing its job when it detects and reports a deviation from the ‘standard’. More on the ‘standard’

and ‘outcasts’ will be argued in the *discrimination and fairness* section. However, when neurodiverse students with, for example, ADHD cannot keep fully focused during the entire exam period and stray their gaze for a long period of time (sometimes paired with distraction from their surroundings), their behavior will be flagged as cheating. Although many diagnosed neurodiverse students notify their examinator about their obstacles in exams, there will always be students who are not diagnosed and are not aware of their neurodiversity themselves. Course administrators who review the report and are not aware of the several ways neurodiverse people act and think, might have the tendency to accept and trust the flag of the algorithm and punish these students when they are neither aware of the neurodiversity of an undiagnosed student. This decision motivated by *confirmation bias*, a form of user bias [67], will be labeled in the training data as a valid example of cheating. The proctoring algorithm then uses machine learning to learn from such biased examples and continues the cycle of further biased accusations. The result could be that the algorithm possibly suffers from training data bias, caused by being fed mislabeled examples of cheating, which are provided by user bias.

4.2.1.2. Discrimination and fairness

It is illegal to discriminate on the grounds of disabilities (including people who are perceived to be disabled) [68]. With regard to ableist bias, the proctoring algorithm discriminates by penalizing neurodiverse students on not behaving ‘normally’.

Neurodiverse people do not prefer to be perceived as sick and disabled, they rather like to be accepted for who they are [69]. However, they are the opposite of neurotypical people – people who think, behave and are perceived as ‘normal’. Consider Danks & London’s formal notion of bias again, wherein they state that bias is a deviation from the standard. The ‘standard’ in this case is ‘normal’ people, and the lack of any deviation for neurotypical people could mean that this implies ‘unbiased’, ‘neutral’ and perhaps ‘fair’. A neurotypical student would be flagged if they behave suspiciously, differing from their own standard, desirable, behavior. This may very well look like cheating if they do not have any good reason to look away from their screen, which is fair and unprejudiced treatment. In contrast, proctoring algorithms perceive the behavior of neurodiverse students as deviations from ‘normal’ behavior from ‘standard’ students, leaving these algorithms biased towards neurodiverse students. Although both neurodiverse and neurotypical students are equally penalized for staring outside of their screen, it is not desirable that they get penalized for equal reasons. Here, take caution that they are treated with counterfactual fairness (see section 4.1). In the eyes of the examiner, this is how fairness is achieved. However, neurodiverse students have several reasons for gazing away, and if they would belong to the group of neurotypical students, it is not likely that their aberrant gaze has a neurologically explainable reason. Both types of student would be flagged, only for the neurodiverse students to be discriminated against on unfair grounds. This is a form of *indirect discrimination* – a form of discrimination wherein

"individuals appear to be treated based on seemingly neutral and non-protected attributes." [41] Yet, while it seems a neutral act for the eye tracking mechanism to report when students gaze away, the data that results from this will reflect implicit effects toward neurodiverse students when their eye movements correlate with their neurodiverse behavior (see [60] and [70]). Eye movements as a proxy for neurodiversity leads to ableist bias.

To achieve fairness towards neurodiverse students, a better notion should be searched, in order to find the right configuration for the automated proctoring algorithm. One suggested notion is *accuracy equity* (see section 4.1), which is achieved when a proctoring algorithm discriminates suspicious behavior and non-suspicious behavior equally well for both groups of students [31], [52]. Neurodiverse students would be flagged on different grounds, which is possible when the algorithm uses two separate classification models. However, this would require a large dataset for both groups separately, which includes the legally protected attribute of disability. Generally, it is a challenging task to collect such data due to privacy limitations and fear of disclosure [66].

4.2.1.3. Mitigation measures

With the goal to achieve fairness through accuracy equity in mind, we can start finding proper mitigation measures to combat training data labelling bias and confirmation bias in the proctoring algorithm. As the first is a consequence of the latter, the first step is to eliminate biased confirmations of flags.

One obvious way is to let the course administrators disable the eye tracking mechanism when determining the settings of cheating behavior. They would have to rely on other aspects, which might give less accurate flags for the true positives. The examinators would have to make a trade-off between a more accurate algorithm and protecting their neurodiverse students. However, this would not guarantee that the proctoring algorithm is free of other implicit, unknown biases. Another way to mitigate confirmation bias is to educate the exam administrators, who make the final judgement, on the signs of neurodiverse behavior. When examinators are aware of a student's diagnosis, they know when a presented flag is a false alarm. Moreover, an undiagnosed student should be treated equally to a diagnosed student when they show similar neurodiverse behavior. So it is important that examinators recognize the signs of neurodiverse behavior, and judge the flags appropriately. On the other hand, in a report on the intersection of bias, disability, and AI, Whittaker et al. [66] ask critically: "When is it appropriate or acceptable to make inferences based on data representing (dis)ability? And, who should be tasked with answering this question?" This is especially problematic when examiners or software owners disclose a student's disability inaccurately and non-consentingly, because it can endanger, for instance, their livelihoods and medical care access [66]. Discrimination towards, for example, autistic people is still common in society [71], so the risk that such information would be unjustly used to exclude and penalize people is high [66]. Thus, we should question:

is it desirable that an examiner judges student's disabilities based on eye tracking data? The discussion of this question lies beyond the scope of this thesis, nevertheless interesting to consider in a full examination of online automated proctoring.

With regard to mitigating bias within the proctoring algorithms themselves, some proctoring services claim to be specially designed to maintain fairness. One such service is Rosalyn.ai. The Rosalyn.ai team states on their website [72]: "We focus as much on the student experience as the efficiency of our system. Diversity, equity, and inclusion are design principles in our remote proctoring system. Performing well for the widest range of body shapes, skin tones and neurotypes is vital to maintaining the integrity of our invigilation as much as the test itself." They believe their proctoring algorithm flags accurately and equitable towards protected groups, because they use a large and ever-expanding training dataset that includes test-takers of all diverse backgrounds [73]. Although they still use eye tracking techniques, using a diverse training dataset to recognize cheating and non-cheating behavior within the neurodiverse group of students, is a step in the right direction towards fairness through accuracy equity.

4.3. What do we need besides a fair algorithm to prevent algorithmic bias?

Many mitigation measures have been proposed, which are often context specific. This section will describe three general ways to mitigate algorithmic bias.

To date, there are few studies that have investigated the association between algorithmic bias and online automated proctoring, and there is little to no research on how to mitigate it. However, from previous work on algorithmic discrimination towards misunderstood people [30], [74]–[76], it is at least reasonable to generally advice a human-centered approach wherein the focus lies on understanding students of all backgrounds. Thus, it is advised to design a decision-making algorithm in collaboration with the protected groups [66]. In other words, foregrounding the experiences of particular groups (e.g. disabled students), rather than forcing them into a painful situation, helps us understand how to overcome bias in data [25]. Moreover, taking an effort to build a developer team with as diverse people as possible is empowering towards protected groups, as the people building an automated decision-making algorithm represent the people who are affected by them.

Another possible mitigation measure that can be applied on most algorithms is by compensating one type of algorithmic bias in one stage with algorithmic bias in another one [12]. This is especially useful when the nature of a occurred bias is known. Danks and London explain that if an algorithm is affected by, for example, training data bias in the sampling process, then it is possible to use a bias in the algorithmic processing stage to offset or correct for the training data bias [12]. Correcting one bias with

another bias is a light measure developers can experiment with, but one must note that developers can be biased as well when they choose the configuration for 'neutralizing' the other bias. One must ask critically: what is their notion of fairness and what is the ideal output of their algorithm when applied to different populations?

The future of ethical AI lies in guidelines and regulations. To date, many governments have attempted to conceptualize lawful regulations on AI ethics, which are often merely a recommendation due to complicated parliamentary processes. For example, in their report named *'Ethics Guidelines for Trustworthy AI'* [77], the European Commission suggested a general 'trustworthy AI assessment list', which is a checklist on legal and ethical requirements to achieve a trustworthy AI. Independently, the Dutch Ministry of the Interior and Kingdom Relations commissioned a team of researchers from Dutch universities together with the Dutch Institute for Humans Rights to compose a step by step handbook on how to mitigate algorithmic bias. The handbook considered technical, legal and organizational criteria in their handbook, which target audience are both the public and private sectors [78]. Such comprehensible legal and ethical frameworks are educational towards developer teams that do not know where to start with developing a fair algorithm.

5. Conclusion

The aim of this thesis was to answer the research question: How can algorithmic bias in automated decision-making algorithms be mitigated to prevent discriminatory decisions? This thesis has shown that algorithmic bias, which we roughly defined as "the worry that an algorithm is, in some sense, not merely a neutral transformer of data or extractor of information" [12], has many different types and sources, and is capable of emerging in all stages of developing an automated decision-making algorithm. Secondly, this thesis found that specifying a notion of fairness is an important step to achieve fairness and prevent discriminatory decisions, but it is a case-specific task. The proctoring case confirmed that it has no straight-forward way to mitigate its bias and achieving fairness requires a developer team to make trade-offs between different values. The findings of this thesis suggest that there is no one-size-fits-all solution to solve every biased algorithm, rather it found that each bias should be examined in its own context. Thus, the research question has no coherent answer, but finding the right mitigation techniques for algorithmic bias depends on the context of the algorithm. However, it seems that collecting representative datasets, developing an algorithm in a diverse team and creating it *with* misunderstood groups could generally help mitigate algorithmic bias. Furthermore, centralized regulations and handbooks might educate developers and bring more awareness to problematic implications of social bias in algorithmic decision-making.

The present thesis has been one of the first attempts to examine online automated proctoring algorithms in the context of algorithmic bias and fairness. Being limited to the amount of previous research, this study lacks an extensive discussion on its mitigation measures. Moreover, this thesis does not claim that it has a definite answer to the ableist fairness problem of automated proctoring. Secondly, it is unfortunate that the thesis did not include more information on different notions of fairness and bias. It was out of the scope of this thesis to provide an extensive list of these notions, which left the lists selective.

Despite its exploratory nature, this thesis offers some insight into how algorithmic bias in automated decision-making algorithms can lead to discriminatory decisions towards protected groups. A natural progression of this work is to analyze what different mitigation techniques are used in which contexts. Considerably more work will need to be done to determine when to use a specific mitigation technique. Moreover, it would be interesting to study discrimination-aware algorithms in future research: *What does it mean for an algorithm to be a non-hypocritical social justice activist?* However, this thesis forms a small stepping stone in such an ever involving field towards more algorithmic fairness.

6. Bibliography

[1] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, “Mitigating Bias in Algorithmic Employment Screening: Evaluating Claims and Practices,” *SSRN Electron. J.*, pp. 469–481, 2019, doi: 10.2139/ssrn.3408010.

[2] M. Richtel, “How Big Data Is Playing Recruiter for Specialized Workers,” *New York Times*, pp. 1–7, Apr. 2013, Accessed: Jul. 16, 2021. [Online]. Available: https://www.nytimes.com/2013/04/28/technology/how-big-data-is-playing-recruiter-for-specialized-workers.html.

[3] M. Bendick and A. P. Nunes, “Developing the Research Basis for Controlling Bias in Hiring,” *J. Soc. Issues*, vol. 68, no. 2, pp. 238–262, 2012, doi: 10.1111/j.1540-4560.2012.01747.x.

[4] C. O’Neil, *Weapons of Math Destruction*. Penguin Books, 2016.

[5] B. Turque, “‘Creative ... motivating’ and fired ,” *The Washington Post*, Mar. 06, 2012. https://www.washingtonpost.com/local/education/creative--motivating-and-fired/2012/02/04/gIQAwzZpvR_story.html (accessed Jul. 17, 2021).

[6] J. M. Tang, M. H. T. de Boer, and S. Vethman, “Technische evaluatie van het AI-algoritme dat de Gemeente Nissewaard inzet voor de opsporing van misbruik en oneigenlijk gebruik van bijstandsuitkeringen,” 2021.

[7] A. Završnik, “Algorithmic justice: Algorithms and big data in criminal justice settings:,” *https://doi-org.proxy.library.uu.nl/10.1177/1477370819876762*, vol. 1, no. 20, Sep. 2019, doi: 10.1177/1477370819876762.

[8] X. F. Aran, T. van Nuenen, J. M. Such, M. Coté, and N. Criado, “Bias and discrimination in AI: A cross-disciplinary perspective,” *arXiv*, pp. 1–6, 2020.

[9] L. Angwin, Julia; Larson, Jeff; Mattu, Surya; Kirchner, “Machine bias.,” 2016. https://www.propublica.org/article/machine-bias-riskassessments-in-criminal-sentencing (accessed Feb. 19, 2021).

[10] M. Zhang, “Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition Software.,” *Forbes*. https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/?sh=d9253c6713d8 (accessed May 13, 2021).

[11] B. Friedman and H. Nissenbaum, “Bias in computer systems,” *Comput. Ethics*, vol. 14, no. 3, pp. 215–232, 2017, doi: 10.4324/9781315259697-23.

[12] D. Danks and A. J. London, “Algorithmic bias in autonomous systems,” *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 0, no. January, pp. 4691–4697, 2017, doi: 10.24963/ijcai.2017/654.

[13] “Gender bias,” *Chemical and Engineering News*, vol. 80, no. 15. p. 12, 2002, doi: 10.1097/00132981-200101000-00001.

[14] S. Barocas and A. D. Selbst, “Big Data ’ S Disparate Impact,” *Calif. Law Rev.*, vol. 104, no. 671, pp. 671–732, 2016.

[15] C. Blom, “Waarom AI niet neutraal is: bias #1 ,” *Innovatief in werk*, Apr. 16, 2019. https://www.innovatiefinwerk.nl/toekomst-van-werk-diversiteit/2019/04/waarom-ai-niet-neutraal-bias-1 (accessed Jul. 15, 2021).

[16] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[17] P. T. Kim, “Data-driven discrimination at work,” *William Mary Law Rev.*, vol. 58, no. 3, pp. 857–937, 2017, Accessed: Jul. 10, 2021. [Online]. Available: https://scholarship.law.wm.edu/wmlr/vol58/iss3/4.

[18] C. Blom, “Waarom AI niet neutraal is: bias #3,” *Innovatief in werk*, May 28, 2019. https://www.innovatiefinwerk.nl/toekomst-van-werk-diversiteit/2019/05/waarom-ai-niet-neutraal-bias-3 (accessed Jul. 16, 2021).

[19] J. Kleinnijenhuis, “Wie wist wat in de toeslagenaffaire? De kluwen van hoofdrolspelers ontward,” *Trouw*, Nov. 14, 2020. https://www.trouw.nl/politiek/wie-wist-wat-in-de-toeslagenaffaire-de-kluwen-van-hoofdrolspelers-ontward~b721c834/?referrer=https%3A%2F%2Fen.wikipedia.org%2F&utm_source=link&utm_medium=social&utm_campaign=shared_earned (accessed Jul. 16, 2021).

[20] J. Henley, “Dutch government faces collapse over child benefits scandal,” *The Guardian*, Jan. 14, 2021.

[21] S. Amaro, “Dutch government resigns after childcare benefits scandal,” *CNBC*, Jan. 15, 2021. https://www.cnbc.com/2021/01/15/dutch-government-resigns-after-childcare-benefits-scandal-.html (accessed Jul. 16, 2021).

[22] “Grondbeginselen rechtstaat geschonden ,” *Tweede Kamer der Staten-Generaal*, Dec. 17, 2020. https://www.tweedekamer.nl/nieuws/persberichten/grondbeginselen-rechtstaat-geschonden (accessed Jul. 16, 2021).

[23] Autoriteit Persoonsgegevens, “Belastingdienst/Toeslagen: De Verwerking van de Nationaliteit van Aanvragers van Kinderopvangtoeslag,” 2020. [Online]. Available: https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/onderzoek_belastingdienst_kinderopvangtoeslag.pdf.

[24] Y. Hofs, “Belastingdienst schuldig aan structurele discriminatie van mensen die toeslagen ontvingen ,” *De Volkskrant*, Jul. 17, 2020. https://www.volkskrant.nl/nieuws-achtergrond/belastingdienst-schuldig-aan-structurele-discriminatie-van-mensen-die-toeslagen-ontvingen~baebefdb/ (accessed Jul. 16, 2021).

[25] Williams, Brooks, and Shmargad, “How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications,” *J. Inf. Policy*, vol. 8, p. 78, 2018, doi: 10.5325/jinfopoli.8.2018.0078.

[26] A. Datta, M. Fredrikson, G. Ko, P. Mardziel, and S. Sen, “Proxy Non-Discrimination in Data-Driven Systems,” 2017. [Online]. Available: http://arxiv.org/abs/1707.08120.

[27] T. Z. Zarsky, “Compensated Surrogacy in the Age of Windsor: Response and Rejoinder:

Understanding Discrimination in the Scored Society," *Washingt. Law Rev.*, vol. 89, p. 1375, 2014, Accessed: Jul. 09, 2021. [Online]. Available: https://digitalcommons.law.uw.edu/wlr/vol89/iss4/10.

[28] L. Zhang, Y. Wu, and X. Wu, "A causal framework for discovering and removing direct and indirect discrimination," in *IJCAI International Joint Conference on Artificial Intelligence*, 2017, pp. 3929–3935, doi: 10.24963/ijcai.2017/549.

[29] T. Baer, *Understand, Manage, and Prevent Algorithmic Bias: A Guide for Business Users and Data Scientists*. 2019.

[30] M. L. Problem and B. Review, "Medicine's Machine Learning Problem | Boston Review," pp. 1–11, 2021, [Online]. Available: https://bostonreview.net/science-nature/rachel-thomas-medicines-machine-learning-problem?fbclid=IwAR2CeA5YKnqP7rCMk4KIJ51wV_H03XGtCllhL5c09pwjABtHoUJ0m9od3m8.

[31] M. Veale and R. Binns, "Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data," *Big Data Soc.*, vol. 4, no. 2, pp. 1–17, 2017, doi: 10.1177/2053951717743530.

[32] N. Altintas, W. Maniram, and J. Veenman, "Discriminatie bij sollicitaties van hogeropgeleide allochtonen?," *Tijdschr. voor Arb.*, vol. 25, no. 1, pp. 83–96, 2009.

[33] Reuters, "Amazon scraps secret AI recruiting tool that showed bias against women | Reuters," *Reuters*, Oct. 10, 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (accessed Jul. 09, 2021).

[34] B. Ruha, *Race after technology: abolitionist tools for the new Jim code*. Medford, MA: Polity.

[35] S. Ghebreab, "Over algoritmen, ongeletterdheid en ongelijkheid." https://www.sennay.net/single-post/2019/02/10/Bevooroordeelde-robots-1 (accessed Jul. 07, 1999).

[36] E. Howarth, "Overrepresentation in criminal justice systems | LSE Undergraduate Political Review," Jan. 25, 2018. https://blogs.lse.ac.uk/lseupr/2018/01/25/overrepresentation-in-criminal-justice-systems/ (accessed Jul. 10, 2021).

[37] B. E. Harcourt, "Against Prediction: Sentencing, Policing, and Punishing in an Actuarial Age," *SSRN Electron. J.*, 2011, doi: 10.2139/ssrn.756945.

[38] S. Corbett-Davies and S. Goel, "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning," 2018, Accessed: Jul. 10, 2021. [Online]. Available: http://arxiv.org/abs/1808.00023.

[39] J. Simon, P. H. Wong, and G. Rieder, "Algorithmic bias and the value sensitive design approach," *Internet Policy Rev.*, vol. 9, no. 4, pp. 1–16, 2020, doi: 10.14763/2020.4.1534.

[40] B. Custers, "Data Dilemmas in the Information Society," *Discrim. Priv. Inf. Soc.*, pp. 1–24, 2013, [Online]. Available: https://ssrn.com/abstract=3047756.

[41] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A Survey on Bias and Fairness in Machine Learning,” 2019, Accessed: Jul. 10, 2021. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[42] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, “Fairness through awareness,” in *ITCS 2012 - Innovations in Theoretical Computer Science Conference*, 2012, pp. 214–226, doi: 10.1145/2090236.2090255.

[43] L. Hu and Y. Chen, “Welfare and Distributional Impacts of Fair Classification,” 2018, [Online]. Available: http://arxiv.org/abs/1807.01134.

[44] P. Rivas, “AI Orthopraxy: Towards a Framework for That Promotes Fairness,” in *2020 IEEE International Symposium on Technology and Society (ISTAS)*, Nov. 2021, pp. 80–84, doi: 10.1109/istas50296.2020.9462167.

[45] S. Vasudevan and K. Kenthapadi, “LiFT: A Scalable Framework for Measuring Fairness in ML Applications,” in *International Conference on Information and Knowledge Management, Proceedings*, Oct. 2020, pp. 2773–2780, doi: 10.1145/3340531.3412705.

[46] L. P. Robert, C. Pierce, L. Marquis, S. Kim, and R. Alahmad, “Designing fair AI for managing employees in organizations: a review, critique, and design agenda,” *Human-Computer Interact.*, vol. 35, no. 5–6, pp. 545–575, Mar. 2020, doi: 10.1080/07370024.2020.1735391.

[47] S. Bird *et al.*, “Fairlearn: A toolkit for assessing and improving fairness in AI,” *Microsoft, Tech. Rep. MSR-TR-2020-32*, pp. 1–6, 2020, Accessed: Jul. 10, 2021. [Online]. Available: https://www.microsoft.com/en-us/research/uploads/prod/2020/05/Fairlearn_WhitePaper-2020-09-22.pdf.

[48] R. K. E. Bellamy *et al.*, “AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias,” *IBM J. Res. Dev.*, vol. 63, no. 4–5, Jul. 2019, doi: 10.1147/JRD.2019.2942287.

[49] J. Kleinberg, S. Mullainathan, and M. Raghavan, “Inherent trade-offs in the fair determination of risk scores,” in *Leibniz International Proceedings in Informatics, LIPIcs*, 2017, vol. 67, doi: 10.4230/LIPIcs.ITCS.2017.43.

[50] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, “Fairness in Criminal Justice Risk Assessments: The State of the Art,” 2017.

[51] M. Kusner, J. Loftus, C. Russell, and R. Silva, “Counterfactual Fairness.” [Online]. Available: https://obamawhitehouse.archives.gov/blog/2016/05/04/big-risks-big-opportunities-intersection-big-data-.

[52] W. Dieterich, C. Mendoza, and T. Brennan, “COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity,” *Perform. COMPAS Risk Scales Broward Cty.*, pp. 1–37, 2016.

[53] S. Subin, “How college students learned new ways to cheat during Covid,” *cnbc.com*, Mar. 21, 2021. https://www.cnbc.com/2021/03/21/how-college-students-learned-new-ways-to-cheat-during-covid-.html (accessed Jul. 11, 2021).

[54] “Proctorio: Securing the integrity of your online assessments.” https://proctorio.com/ (accessed Jul. 13, 2021).

[55] “Respondus.” https://web.respondus.com/ (accessed Jul. 13, 2021).

[56] “Honorlock On-Demand Online Proctoring Service,” 2021. https://honorlock.com/ (accessed Jul. 13, 2021).

[57] “ProctorU - The Leading Proctoring Solution for Online Exams,” 2021. https://www.proctoru.com/ (accessed Jul. 13, 2021).

[58] S. Swauger, “Software that monitors students during tests perpetuates inequality and violates their privacy,” *MIT Technology Review*, Aug. 07, 2020.

[59] M. Harris, “Viral TikTok: Student Fails Exam After AI Software Flags for Cheating,” *Insider*, Oct. 04, 2020.

[60] M. Retta, “Exam Surveillance Tools Monitor, Record Students During Tests | Teen Vogue,” Oct. 26, 2020. https://www.teenvogue.com/story/exam-surveillance-tools-remote-learning (accessed Jul. 11, 2021).

[61] A. Khan, “The @ExamSoft software can’t ‘recognize’ me due to ‘poor lighting’ even though I’m sitting in a well lit room. Starting to think it has nothing to do with lighting. Pretty sure we all predicted their facial recognition software wouldn’t work for people of,” *Twitter*. Oct. 08, 2021, Accessed: Jul. 11, 2021. [Online]. Available: https://twitter.com/uhreeb/status/1304451031066083331.

[62] M. Chin, “ExamSoft’s proctoring software has a face-detection problem - The Verge,” *The Verge*, Jan. 05, 2021. https://www.theverge.com/2021/1/5/22215727/examsoft-online-exams-testing-facial-recognition-report (accessed Jul. 12, 2021).

[63] K. Johnson, “ExamSoft’s remote bar exam sparks privacy and facial recognition concerns | VentureBeat,” *Venture Beat*, Sep. 29, 2020. https://venturebeat.com/2020/09/29/examsofts-remote-bar-exam-sparks-privacy-and-facial-recognition-concerns/ (accessed Jul. 12, 2021).

[64] B. Stewart, “Online exam monitoring can invade privacy and erode trust at universities,” *The conversation*, Dec. 03, 2020. https://theconversation.com/online-exam-monitoring-can-invade-privacy-and-erode-trust-at-universities-149335 (accessed Jul. 17, 2021).

[65] R. Heilweil and S. Morrison, “Online proctoring services pose privacy concerns for remote learning ,” *Vox*, Dec. 18, 2020. https://www.vox.com/recode/22175021/school-cheating-student-privacy-remote-learning (accessed Jul. 17, 2021).

[66] M. Whittaker *et al.*, “Disability, Bias, and AI,” New York, Nov. 2019.

[67] “What is Bias in Machine Learning & Deep Learning?,” *Foresee Medical*, Apr. 20, 2020. https://www.foreseemed.com/blog/bias-in-machine-learning (accessed Jul. 12, 2021).

[68] Nederlandse Grondwet, “Artikel 1: Gelijke behandeling en discriminatieverbod .” https://www.denederlandsegrondwet.nl/id/vgrnb2er8avw/artikel_1_gelijke_behandeling_en?v=1&ctx=vgrnb2er8avw (accessed Jul. 12, 2021).

[69] “Vijf vragen over Neurodiversity Pride Day,” *Nederlandse Vereniging voor Autisme*, Jun. 15, 2020. https://www.autisme.nl/2020/06/15/vijf-vragen-over-neurodiversity-pride-day/ (accessed Jul. 12, 2021).

[70] L. M. Schmitt, E. H. Cook, J. A. Sweeney, and M. W. Mosconi, “Saccadic eye movement abnormalities in autism spectrum disorder indicate dysfunctions in cerebellum and brainstem,” *Mol. Autism*, vol. 5, no. 1, 2014, doi: 10.1186/2040-2392-5-47.

[71] “Disability discrimination in further and higher education,” *National Autistic Society*. https://www.autism.org.uk/advice-and-guidance/topics/education/resolving-differences/disability-discrimination-in-further-and-higher-ed (accessed Jul. 17, 2021).

[72] “About Rosalyn,” *Rosalyn AI*. https://www.rosalyn.ai/about (accessed Jul. 16, 2021).

[73] “Your Online Exam Experience Matters: How 4 Students Rate the Most Popular Platforms,” *Rosalyn AI*. https://www.rosalyn.ai/blog/your-online-exam-experience-matters-how-4-students-rate-the-most-popular-platforms-ros (accessed Jul. 17, 2021).

[74] N. U. Din *et al.*, “Age and gender variations in cancer diagnostic intervals in 15 cancers: Analysis of data from the UK clinical practice research datalink,” *PLoS One*, vol. 10, no. 5, 2015, doi: 10.1371/journal.pone.0127717.

[75] G. Lyratzopoulos, G. A. Abel, S. McPhail, R. D. Neal, and G. P. Rubin, “Gender inequalities in the promptness of diagnosis of bladder and renal cancer after symptomatic presentation: Evidence from secondary analysis of an English primary care audit survey,” *BMJ Open*, vol. 3, no. 6, p. e002861, Jun. 2013, doi: 10.1136/bmjopen-2013-002861.

[76] C. Criado-Perez, *Invisible women : data bias in a world designed for men*. .

[77] “Ethics guidelines for trustworthy AI,” 2019.

[78] B. van der Sloot, E. Keymolen, M. Noorman, H. Weerts, Y. Wagensveld, and B. Visser, “Non-discrimination by design.” pp. 1–69, 2021.

[79] S. Coghlan, T. Miller, and J. Paterson, “Good proctor or ‘Big Brother’? AI Ethics and Online Exam Supervision Technologies,” 2020, Accessed: Jul. 12, 2021. [Online]. Available: http://arxiv.org/abs/2011.07647.