UPPSALA
UNIVERSITET

UPTEC STS 19033
Examensarbete 30 hp
Juni 2019

Mitigating algorithmic bias in Artificial Intelligence systems

Johanna Fyrvald

UPPSALA
UNIVERSITET

Teknisk- naturvetenskaplig fakultet
UTH-enheten

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

Abstract

Mitigating algorithmic bias in Artificial Intelligence systems

Johanna Fyrvald

Artificial Intelligence (AI) systems are increasingly used in society to make decisions that can have direct implications on human lives; credit risk assessments, employment decisions and criminal suspects predictions. As public attention has been drawn towards examples of discriminating and biased AI systems, concerns have been raised about the fairness of these systems. Face recognition systems, in particular, are often trained on non-diverse data sets where certain groups often are underrepresented in the data. The focus of this thesis is to provide insights regarding different aspects that are important to consider in order to mitigate algorithmic bias as well as to investigate the practical implications of bias in AI systems. To fulfil this objective, qualitative interviews with academics and practitioners with different roles in the field of AI and a quantitative online survey is conducted. A practical scenario covering face recognition and gender bias is also applied in order to understand how people reason about this issue in a practical context. The main conclusion of the study is that despite high levels of awareness and understanding about challenges and technical solutions, the academics and practitioners showed little or no awareness of legal aspects regarding bias in AI systems. The implication of this finding is that AI can be seen as a disruptive technology, where organizations tend to develop their own mitigation tools and frameworks as well as use their own moral judgement and understanding of the area instead of turning to legal authorities.

Handledare: Patrick Couch
Ämnesgranskare: David Sumpter
Examinator: Elísabet Andrésdóttir
ISSN: 1650-8319, UPTEC STS 19033

Sammanfattning

Artificiell intelligens (AI) appliceras i dagens samhälle i allt större utsträckning. AI-system har en enorm potential att lösa storskaliga problem och appliceras redan i en mängd olika beslutssammanhang som ofta har en direkt påverkan på människan. AI-system används exempelvis för att besluta om vilken behandling patienter ska få, vem som blir tilldelad ett banklån eller vem som blir föreslagen som kandidat till en viss roll inom en organisation. Eftersom att dessa AI-system kan ha en så stor inverkan på våra liv och ta beslut som avgör människans framtid är det av stor vikt att dessa system är rättvisa, inkluderande och att algoritmerna som AI-systemen använder sig av för att ta beslut inte innehåller bias eller ger ett snedvridet och diskriminerande resultat.

Bias inom AI-system är ett komplext, socioteknisk problem, där en stor utmaning ligger i att AI-system ofta tränas på historiska data, som speglar de normer och beslut som har tagits i samhället under en längre period. Vissa grupper i samhället har exempelvis historiskt sett blivit diskriminerande på grund av kön, etnisk tillhörighet, ålder etc. och algoritmisk snedvridning kan därmed uppstå när den data som AI-systemen tränas på inte är tillräckligt representativ för alla grupper i samhället. När denna data som redan är icke-diversifierad matas in i systemen replikeras samt amplifieras dessa mönster ofta i utfallet och i besluten som AI-systemen tar. Det finns flera exempel på denna problematik från verkligheten och en viss typ av AI-system, ansiktsigenkänningssystem, har visat sig ofta vara tränade på icke-diversifierade data. Ett exempel på detta är när Google Photos ansiktsigenkänningssystem år 2015 av misstag klassificerade två mörkhyade personer som Gorillor, vilket troligtvis berodde på att AI-systemet inte tränats tillräckligt mycket på alla typer av utseenden och grupper i samhället.

Syftet med denna studie är att undersöka detta område genom att utreda olika aspekter som är viktiga att ta i beaktning för att minska algoritmisk bias och diskriminering i AI-system, samt undersöka de praktiska implikationerna av algoritmisk snedvridning. För att uppfylla studiens syfte genomfördes sju kvalitativa intervjuer med personer som är aktiva inom området AI från både akademi och näringsliv. Dessutom genomfördes en kvantitativ enkätstudie som besvarades av 303 privatpersoner. Ett praktiskt scenario som behandlade bias med avseende på kön i ett AI-system för ansiktsigenkänning användes i både de kvalitativa intervjuerna samt i den kvantitativa enkätstudien. Syftet med det praktiska scenariot var att undersöka hur personer med olika bakgrund resonerar kring vad som är rättvist i AI-system samt huruvida de ställer olika aspekter såsom bias, rättvishet, säkerhet i jämförelse med prestandan och kvalitén på systemet mot varandra. Den data som samlades in under intervjuerna sammanställdes och analyserades med tematisk analys. Sju teman identifierades och användes för att beskriva intervjupersonernas argument i syfte att sedan kunna dra slutsatser utifrån dem.

Studien visade huvudsakligen att representanterna från näringslivet hade en djupare medvetenhet och förståelse för potentiella risker och utmaningar med bias i AI-system

än akademikerna. Representanterna från näringslivet har även i större utsträckning börjat implementera samt arbeta mer praktiskt med tekniska lösningar för att minska algoritmisk snedvridning i deras AI-system än akademikerna. Studien visade även att, trots en hög medvetenhet kring ämnet, hänvisade ytterst få av intervjupersonerna till legala aspekter när de resonerade kring sina svar. Representanterna från näringslivet utvecklar även till stor del sina egna tekniska lösningar för att få bukt på problemet med snedvridning inom AI-system. Det finns därmed en risk att AI-system kan ses som en disruptiv teknologi, vilket innebär att teknikens implikationer på samhället är ovisa och osäkra. Då dessa tekniker tar över marknader i en snabb takt finns en risk att de legala aspekterna undangöms för att företagens egna moraliska bedömningar och förståelse för ämnet och tekniken prioriteras i första hand.

Förord

Detta examensarbete utfördes som det avslutande momentet inom civilingenjörsutbildningen System i Teknik och samhälle (STS) vid Uppsala Universitet. Arbetet har skett i samarbete med företaget IBM Svenska AB, där jag vill tacka min handledare Patrick Couch för vägledning och stöd i form av framförallt idéer, engagemang, förslag på olika infallsvinklar inom uppsatsämnet samt värdefulla kontakter som bidragit till den huvudsakliga datainsamlingen till examensarbetet.

Jag vill även tacka min ämnesgranskare David Sumpter, vid den Matematiska institutionen vid Uppsala Universitet, som har varit en stor hjälp under examensarbetets gång genom regelbunden uppföljning, förslag på tillvägagångssätt, konstruktiv feedback och ett brinnande intresse för det valda ämnesområdet.

Johanna Fyrvald
Uppsala, juni 2019

Table of contents

1. Introduction

Artificial intelligence holds an enormous potential of transforming businesses, solving large scale problems and making critical decisions. AI systems are being increasingly used and widespread in society as well as make and support an enormous amount of decisions that directly affects people's lives; such as credit scoring solutions, crime prediction methods and recruitment tools (IBM Policy, 2018). Since the outcome of the decisions that the AI systems make can have direct implications on the life of humans, it is critical that these systems adopt a responsible behaviour that is fair, unbiased and non-discriminatory with specific regard to sensitive features such as gender, ethnical background and age (Kamishima et al., 2011).

A challenge with AI systems is that the inherent structures of society as well as the history of discrimination against minority groups can be represented in the data that AI systems are trained on. This implies that if societal bias or prejudices are present in the data set, these features can get replicated or even amplified in the outcome of the decisions, hence AI systems are only as good as the data that they are trained on (Howard and Borenstein, 2018). Bias can also be hard to detect and identify since humans may not be consciously aware of their existence in the data, which is referred to as implicit bias (Brownstein, 2016). Solutions are therefore needed to integrate moral, societal and legal values along with the technical progress and design processes of AI systems (Dignum, 2017).

Algorithmic bias can emerge when the data distribution that an AI system is trained on not is representative or diverse enough of the situation one wants to model and reason about (Srivastava and Rossi, 2018). There are several real-world examples when AI systems has been trained on data which is not inclusive or diverse enough to represent the entire population in a fair way. The giant tech company Amazon's automated recruiting engine was for example penalizing resumes including the word "women's", since the AI system was trained on historical data over a 10-year period of time where male applicants were dominant and therefore seen as preferable for certain positions (Dastin, 2018).

1.1 Background and previous work

1.1.1 Algorithmic bias in AI systems

The problematic issue with bias in AI systems has received increased attention in media and in public technology related discussions (Danks and London, 2017). One of the cases that has brought a lot of attention regarding the issue of racial bias, is an algorithm called COMPAS which is used is the U.S. criminal justice system to perform risk assessments in regard to a criminal defendant's likelihood of re-offending (Angwin et al., 2016). When ProPublica, an American non-profit newsroom, analysed the efficacy

of COMPAS the results showed that black defendants were twice as likely as white defendants to be misclassified as a higher risk of reoffending, while white defendants were more likely to be incorrectly judged as a low risk of reoffending. Hence, the algorithm's predictions seemed to favour white defendants over black defendants through underpredicting the risk for white defendants to reoffend and overpredicting the corresponding risk for black defendants (Larson et al., 2016). However, when ProPublica presented these results, the company who created the COMPAS algorithm, Northpointe, responded in form of a research paper where they stated that ProPublica had based their results on the wrong classification statistics and therefore had misunderstood the meaning of an algorithm making an error. Northpointe also claimed that their algorithm was equally well calibrated and tested as other similar algorithms used in society. Northpointe therefore strongly rejected that COMPAS is racially biased against black people (Dieterich et al., 2016). This underlines how complex the issue of algorithmic bias is and that the moral judgement of the developers of AI systems and algorithms can conflict with legal norms.

Although the data used by COMPAS do not use an individual's race as a feature when doing predictions, other combined data can be correlated to a person's race which can lead to bias against black people in the predictions (Dressel and Farid, 2018). Applying simplistic solutions such as removing sensitive features from a data set which can cause bias in the system, is therefore not a viable solution for mitigation of algorithmic bias in AI systems (Kamishima et al., 2011).

1.1.2 Mitigating algorithmic bias

In order to mitigate bias, research have been conducted by many leading global technology companies. IBM is one of the organizations that are conducting extensive research with the aim of accelerating the area of fairness in AI systems. A technical solution that IBM has developed for this purpose is called AI OpenScale. The objective with this tool is to ensure that AI systems are performing well and accurate as well as produce fair outcomes. The platform provides insights regarding three main parameters; fairness, accuracy and explainability of an AI system. The solution automatically detects unfair or inaccurate results and detect, mitigate as well as explains bias and model outcomes. Additionally, AI OpenScale allow the user to trace back a decision-making process in order to explain what factors that influenced the decision and how the different factors might have changed the outcome (Howard and Howard, 2018). Another tool which IBM has developed is an open source Python toolkit for algorithmic fairness called AI Fairness 360. The main objective with the toolkit is to facilitate research within the area of AI fairness and to detect, understand and mitigate unwanted algorithmic bias (Bellamy et al., 2018).

A specific area within AI that is receiving increased attention is face recognition systems. However, most existing face recognition tools are trained and tested on non-

diverse data sets and have an underrepresentation in the training data for certain groups. An example of this issue is when Google Photos face recognition tool in 2015, mistakenly classified two dark-skinned people as “gorillas” (Garcia, 2016). In order to prevent further discriminating cases similar to this, IBM has developed a data set called “Diversity in Faces Data set”, which is an open data set that consists of one million images of faces. The data set is designed with the aim of promoting the use of inclusive data and to advance the study of fairness and mitigation of bias in face recognition technology (Merler et al., 2019).

In addition to technical solutions for mitigation of bias in AI systems, non-technical aspects are important to consider as well. The global IT software and service company Tieto, has for example introduced their own AI ethical Guidelines with the purpose of reinforcing their commitment to a responsible development of AI systems. Tieto also shows the importance of an increased awareness and understanding for the area, through introducing AI ethics certifications for their employees. Additionally, the organizations are planning on establishing new AI ethics roles, such as transparency engineers and AI trainers that can teach AI ethics to the organizations AI solutions (Tieto, 2018).

1.1.3 AI and face recognition as disruptive technologies

Disruptive technologies, which first was introduced by Christensen (1997) are technologies that challenges incumbent business areas and target new market segments (Christensen, 1997). A disruptive technology is developed in a fast pace and is often forecasted to revolutionize the lives of humans. The impact and implications on society that a disruptive technology will lead to is often unknown and uncertain. AI and Face recognition, in particular, can be seen as disruptive technologies since it is reshaping society in several ways and the consequences can lead to uncertain implications. Therefore, disruptive technologies often induce a need for a reassessment of the existing legal frameworks and regulations and if needed, changes in the law (Kolacz and Quintavalla, 2019).

However, there is a risk that disruptive technologies ignore legal frameworks when entering new markets or that organizations try to identify loopholes in the law. This idea makes it possible for organizations to distribute a technology quickly and spread it amongst several markets before law and legal authorities are catching up (Isaac and Davis, 2014). An example of this issue is the revolutionary technology company, Uber, which is an on-demand taxi service based on a smartphone app. Uber has disrupted the entire taxi industry through their low-fixed cost model and fast service, amongst other factors. The company managed to put themselves in a position of a legal void, since the organisation chose to operate in terms of a technology company instead of a transportation firm. Therefore, the company has been able to find creep-holes in the law and regulations of different markets and can therefore avoid costly regulations and employer responsibilities (Isaac and Davis, 2014). This issue shows the risk that

disruptive technologies can override legal aspects and as soon as the users in different markets have adopted the technology, it's more complex to go back and change the regulatory frameworks.

1.1.4 Normative and descriptive work in algorithmic fairness

The concerns around algorithmic bias in AI systems have led to increased research and recent work covering different approaches to mitigate bias. However, the focus of most existing work is to find and mitigate discrimination in regard to fairness in decision-making processes. The majority of the studies of algorithmic fairness is therefore normative, meaning that it is focused on how non-discriminatory decisions are supposed to be made. However, in this thesis, a complementary descriptive approach towards fair decision making is applied using a practical scenario (Grgic-Hlaca et al., 2018). The descriptive approach is applied through focusing on a specific context with the aim of understanding how people reason about unfair bias and discrimination in AI systems. The goal with the additional descriptive approach is to investigate the moral reasoning behind arguments and perceptions that people working with AI and individuals apply in a practical scenario. Further on, discussion regarding how the findings from the empirical studies can be used in order to mitigate unfair bias in AI systems is conducted.

1.2 Research objective

The objective with this thesis is to provide insights regarding different aspects that are important to consider in order to mitigate algorithmic bias as well as to investigate the practical implications of bias in AI systems. To address this objective the current awareness and understanding of the challenges and potential risks with introduction of unfair bias in AI models will be investigated and the different methods that can be applied in order to mitigate algorithmic bias. The practical implications are investigated through a practical scenario that is applied in order to project the ethical dilemmas around algorithmic bias in a potential real-world application and to understand how organizations can handle the problematics around algorithmic bias practically. In order to fulfil the objective of the study, the following research questions will be examined:

- To what extent are organizations aware of and understand the potential risks and challenges with algorithmic bias in AI systems?
- What technical and non-technical solutions can be implemented in order to mitigate algorithmic bias in AI systems?
- How do people in different organizations respond to a practical scenario related to algorithmic bias presented to them?
- What are the differences and similarities in how people in different organizations and individuals respond to an online survey covering a practical scenario?

2. Methods

In the following section, the methods that are applied in this study is described. The choice of quantitative and qualitative research methodology is motivated, followed by the design of the questionnaire and guidelines for the practical scenario. Thereafter, the study subjects that were interviewed in the study are described and the choice of academics and practitioners is motivated. Finally, the method of analysis of the gathered data is presented.

2.1 Research methodology

The main research methodology that has been applied in this study is qualitative. A qualitative method is appropriate to use when a deeper understanding of a specific phenomenon or research area will be investigated. Since the research area that is investigated in this study is in a relatively early stage the study is of an explorative character with the aim of covering different aspects of mitigation of algorithmic bias. A qualitative method that is explorative has relatively low restrictions and high flexibility and it allows for processes to be modified along with the observations, which made this approach suitable for this study. Interviews are a very common data collection method used in qualitative studies. The primary data collection method for this study was semi-structured interviews. Semi-structured interviews are characterized by a number of questions that are prepared in advance, but it also opens up for follow-up questions in order to gather more details regarding a specific response or area that the respondent seem to be very familiar with (Graziano and Raulin, 2013). In addition to the qualitative methodology a complementary quantitative online survey was conducted. The aim with the quantitative study was to gather a statistical result of what a group of individuals from the general public perceive as fair and investigate what aspects they believe are important to consider when developing AI systems in order to mitigate the risk of algorithmic bias in the systems.

2.2 Questionnaire

In order to investigate the first two research questions regarding awareness and understanding of challenges and potential risks with algorithmic bias in AI systems as well as technical and non-technical methods to mitigate those, a questionnaire was formed. The questionnaire consisted of eight questions regarding the respondent's own perception and experience of bias in AI systems and how the respondent's organization approaches this research area. The questions covered relevant areas related to the research questions and different aspects of algorithmic bias in AI systems. Aspects such as potential risks with bias in AI systems, stakeholder engagement, practical steps for mitigating bias, transparency in the decision-making process of AI systems as well as diversity in development teams and the data sets used to train AI systems on were

covered, see Appendix A. All the respondents were asked very similar questions initially in order to make it possible to compare their answers. The follow-up questions varied depending of the expertise and knowledge of the respondent.

2.3 Scenario

In the second part of the qualitative interviews, the participants were presented questions related to a made-up practical scenario in the context of a face recognition system and gender bias. The purpose with the scenario was to let the interviewees reason out loud about questions related to a potential real-world application in order to learn how they reasoned their way to decide whether an AI system is unfair, or gender biased against a specific group and the reasons why. The different questions provided various information and related numbers in order to see whether the participant's opinion changed depending on what information was provided about the system. The scenario was also used in the quantitative part of the study in the form of an online survey. The objective with the use of the scenario was to investigate the last two research questions regarding the responses to a practical scenario related to algorithmic bias in AI systems as well as differences and similarities in how the interviewees responded to a practical scenario compared to how they participants in the online survey responded. The thoughts behind the design of each question in the scenario and the guidelines regarding how the scenario was ought to be solved are described below.

2.3.1 Scenario design & guidelines

The scenario was designed with a potential real-world application in mind, in form of an AI system in order to make the scenario as relevant and realistic as possible. Therefore, an AI system for face recognition that is supposed to automatically open the door for employees in an organization was used. The description of the scenario, see figure 1, was initially presented to the participants followed by five related questions presented one by one. There are no definite right or wrong answers to the questions in the scenario, but the guidelines provide the reasoning behind the design of the questions and suggest answers to the questions regarding how the scenario could be solved.

Description of Scenario: AI system for face recognition

An AI system has been developed to automatically open the door of a building for employees using face recognition. The system uses a database of all 500 employees, 100 of whom are women and 400 are men, plus a test data set of 500 non-employees (also 100 women and 400 men).

The system is required to have an error rate (both false positives and false negatives) of less than or equal to 10% to reach an acceptable operational standard. Consider the following questions about the system. There are no definitive "wrong" or "right" answers, I want to understand your thinking about these problems.

Figure 1. The description of the scenario presented to the interviewees.

The first question related to the scenario provides information about the false positives rates for both men and women, see figure 2. The false positive rate describes the number of non-employees that the system mistakenly lets in. In this question the system mistakenly opens the door for 10 out of 100 female non-employees from the test data set which is equal to a 10% false positive rate, while the corresponding rate for male non-employees is 2,5%. The guideline for solving this question is that there is not enough information presented in the question in order to determine whether the system is gender biased or not. The reason for this is since the question only presents information about how the system behaves for the non-employees, but there is no information provided about how the system treats the actual employees. However, one could argue that women are unfairly treated in this case since the system recognizes less female non-employees and it therefore has a higher error rate for women than for men. On the other hand, one could also argue that men are being unfairly treated since more women than men are being let in by the system. In order to see whether the respondents would suggest any further tests that could be applied to determine whether the system is gender biased or not, a follow up question regarding what further tests that could be carried out was designed. The guideline to the follow-up question is to run tests with the aim of finding out the false negative rate, since that rate cannot be calculated from the information provided. The false negative rate reveals how many females versus males who actually are employees that the system doesn't permit entry to.

Question 1

From the test data set of non-employees, the AI system mistakenly opens the door for 10 out of 100 female non-employees and 10 out of 400 male non-employees.

- *Do you think that the system is gender biased? If so, is it men or women who are unfairly treated and why?*
- *What further tests would you ask your team to carry out?*

Figure 2. Question 1: Related to the false positive rates from the test data set of non-employees and gender bias and a follow-up question regarding further tests.

In question 2, information about further tests covering the actual employees is presented. The AI system now permits entry to 105 women in total, 95 employees and 10 non-employees, which gives the same false positive rate of 10% for females, as in the previous question. For males, the corresponding numbers are a total of 400 men, where 390 employees and 10 non-employees are permitted entry. The false positive rate for men remains the same (2,5%) as in the previous question, see figure 3. The further tests that are carried out in this question also provide the numbers for calculating the false negative rate, which for women is 5 out of 100 which equals 5%, and for men 10 out of 400 which equals 2,5%. The guidelines for solving this question could therefore involve calculations of these rates and especially evaluation of the difference in performance regarding between men and women in the false negative rates. With this information it is clearer than in the previous question that the system could be gender biased against women, due to the fact that more women who are employees that should have been let in by the system gets denied permission to their workplace.

Question 2

Further tests show that the AI system permits entry to 105 women, 95 of whom are employees and 10 who are not employees. It also permits entry to 400 men, 390 of whom are employees and 10 of whom are not employees.

- *Do you think that the system is gender biased?*
- *If so, is it men or women who are unfairly treated and why?*

Figure 3. Question 2: Includes information about further tests and the false negative rates related to gender bias.

In question 3, the false negative rates for both genders are explicitly written with the aim of providing more clarity about the differences, percentage wise, see figure 4. At this stage it should be clear that the system is gender biased against women, since more women who are employees are denied entry by the system and one could therefore

argue that they are unfairly treated. However, both of the rates are still lower than the 10% error rate that is stated in the initial description of the system, which is required for an acceptable operational standard.

Question 3

Your technical team reminds you that out of the 100 female employees, 5 are not permitted entry by the system (5% false negative rate). Of the 400 male employees, 10 are not permitted entry by the system (2.5% false negative rate).

- *Now do you think that the system is gender biased? against men or women?*

Figure 4. Question 3: Further clarification regarding the false negative rates related to gender bias.

Moving forward in the scenario, question 4 was designed in order to learn how the participants weighted the parameters of how fair a system is weighted against the performance. This was done through comparing the initial system to an "improved" system, see figure 5. In this question, the technical team has made an improvement of the system and presents a new system where the error rates are exactly the same for men and women, 10% false positive error rate in both cases and 5% false negative error rate. In this case, the system could be seen as fairer in one way, but what has actually happened is that the accuracy has only become worse for men and the error rates are remaining the same as in the initial system for women.

The ethical dilemma in this question is therefore whether this "improved" system is better than the initial system or not. Discussion regarding whether it can be justified to make the system perform worse for men in order to achieve equal error rates or not is expected from the participants. The guideline regarding this question is that, solving the problem through increasing the error rates for men in order to make the system fairer is not a viable solution. Alternative solutions such as increasing the data set for both men and women in order to improve the accuracy for both genders could also be discussed. Depending on what system the participant choses as the preferred system, follow-up questions were applied in order to make the participants reflect over their choice. If the participant chose the initial system, a follow-up question regarding how they would justify that they chose a system which they knew treats women unfairly was asked. On the other hand, if the participants preferred the "improved" system, a follow-up question regarding how they would justify that the system performs worse for men on purpose in order to make it fairer, was asked. In that case, discussion regarding whether the system then could be seen as biased against men instead was held.

Question 4

After making some revisions your team announces it has improved the system so that the errors are the same for men and women. Now the system incorrectly opens the door for 10 out of 100 women and 40 out of 400 men (10% false positive error in both cases). It correctly permits entry to 95 out of 100 women and 380 out of 400 men (5% false negative error in both cases).

- *Is this new system better than the one used in questions 1-3? Discuss.*

Figure 5. Question 4: An improved system with the same error rates for men and women are presented. Discussion regarding whether the participants prefer the initial system, or the improved system was held.

In the final question, the participants were supposed to assume that men are 10 times as likely to illegally enter office premises and commit crimes, see figure 6. This question was designed with the purpose of learning how the participants would reason and decide if they would include this information when developing the system or not. Initially, one might think that this information should be included when developing the system in order to reduce the error rates for men, since it can be considered a safety critical aspect if a man that is more likely to commit a crime gets permitted access into the building. However, further reflection and consideration over the dilemma was covered by a follow-up question regarding if the respondents would treat a man stricter if a man and a woman was standing outside the building wanting to get in. The participants then needed to reflect on whether they would change their mind or stick to their initial decision. It could be considered as discriminating against men to generalize like that and treat them stricter than women because of this information, it might even not be justified due to legal aspects. The guideline regarding this question would therefore be not to use this information when developing the system or treat men stricter, since it would not be considered fair or might not even be legal.

Question 5

Now assume that men are 10 times as likely to illegally enter office premises and commit crimes.

- *How would you use this information when developing the system? Discuss in context of the AI system we have developed.*

Figure 6. Question 5: The participants were expected to take a stand regarding how they would use this new information in the context of the developed AI system.

2.3.2 Online survey

The quantitative survey was designed in the form of replicating the scenario that was presented to the interviewees in the qualitative study. The purpose with the quantitative survey was to collect the individuals and potential end users of AI systems opinion regarding what they perceive as fair as well as what aspects of the system that they would prioritize when designing an AI system. The quantitative part added statistical results and further understanding of different aspects to consider in order to mitigate bias when developing AI systems and provided a wider range of potential answers that could be used to compare the answers from the qualitative interviews with the responses from the online survey. The questions in the survey were refined in order for the respondents to choose between several alternatives in their responses. The respondents were also given an opportunity to explain why they chose a specific alternative. The full quantitative survey can be found in Appendix B. The survey was created with Google Forms and primarily distributed through Twitter as well as shared on other social media platforms and with friends and family.

2.4 Study subjects

The study subjects for this study consist of a mix between academics as well as practitioners which are active within the field of Artificial intelligence. Since the main research method for this study is qualitative, a few persons have been carefully chosen and studied at a deeper, more detailed level. The respondents were chosen with the aim of finding persons who could enhance the understanding of the studied phenomena with their knowledge and expertise and to investigate their opinions in accordance with the objective of the study and the research questions. The selection of respondents was not specified from the beginning and the selection developed over time, which is a common methodology in qualitative studies (Miles et al., 2014). The study subjects for the qualitative interviews were chosen based on a combination of recommendations and contacts provided by my supervisor, reviewer and personal network. The participants chosen represented both academia and practitioners from the industry. This choice was made in order to investigate whether respondents from academia compared to industry representatives would reason similarly or if any clear differences in their responds would appear. Since both researchers and academics as well as representatives from the industry play a major role in the development process of fair and inclusive AI systems, respondents from both of these groups were included. The interviewees gave consent to participate with their name and role, except for one participant that wanted to remain anonymous, see table 1, and all the interviews were held in English.

Table 1: Summary of study subjects and interviews

Interviewee	Gender	Organization	Subject category	Date/ length	Type of interview/place
Devdatt Dubhashi, Professor - Data Science & AI	Male	Chalmers University	Academic	5/3-19, 40 min	Personal, Lindholmen Science Park
Staffan Truvé, CTO & Co - founder	Male	Recorded Future	Practitioner	18/3-19, 40 min	Video interview
Jana Tumova, Assistant professor - Robotics	Female	Royal Institute of Technology	Academic	26/3-19, 35 min	Personal, Royal Institute of Technology
Christian Guttmann, VP & Global Head of AI	Male	Tieto	Practitioner	27/3-19, 30 min	Phone interview
Nasim Farahini, CTO - AI, IoT & Cloud	Female	Qamcom	Practitioner	4/4-19, 75 min	Personal, Kista
Anonymous, Software Engineer - Credit risk assessment	Female	IT-company (Anonymous)	Practitioner	19/4-19, 40 min	Written interview
Andreas Theodorou, Researcher - Responsible AI	Male	Umeå University	Academic	29/4-19, 40 min	Written interview

2.4.1 Academics

The academics that were chosen to participate in the qualitative interviews represented three major universities in Sweden; Chalmers University, The Royal Institute of Technology and Umeå University. All the academic respondents were researchers within Artificial Intelligence related areas such as AI, Data Science and Robotics. Devdatt Dubhashi, professor in data science and AI at Chalmers University with particular focus on design and development of randomizing algorithms and Machine Learning for Big data, was specifically chosen since he is part of the operating team of Chalmers AI Research Centre, Chair. Chair is a recent initiative were the focus is to enhance Chalmers expertise and excellence within the area of AI and bring together industry representatives, academics and governmental institutions. The Chair initiative also consists of an AI-ethics group, that is supposed to ensure that ethical aspects are permeated throughout the research and performed activities (Chalmers, 2019).

Jana Tumova, assistant professor in Robotics at the Royal institute of technology was chosen as a respondent in order to add the academic perspective from a different area within AI, Robotics. Robotics is a highly relevant area within AI and since robotics systems are interacting with humans on a daily basis in form of smart home systems, work related robots, chatbots etc., it is critical that these systems are unbiased and treat the humans they interact with in a fair way. Andreas Theodorou, Researcher in AI, was chosen as a respondent since he focuses on the design and application of intelligent systems and its effects on humans and society. Theodorou conducts his research at the Responsible AI Research Group at Umeå University and he is currently working on designing guidelines in order to integrate the socio-economic, legal and ethical considerations which arise from implementing AI in society. Theodorou's research interest and current work is therefore highly relevant for the purpose of the study and he could also provide unique insights regarding the intersection between AI and its ethical implications in society.

2.4.2 Practitioners

The practitioners who were selected as respondents represent both global companies with a large-scale adoption of AI development and smaller firms that are in the early stages of their AI research and deployment. This selection was done in order to cover a number of organizations of various sizes and focus areas within the field of AI with the aim of investigating differences and similarities in their perspectives about algorithmic bias. Staffan Truvé, Co-founder and CTO of Recorded Future, was chosen as a study subject since Recorded Future is a strongly data-driven organization which collects and analyses vast amounts of data in order to deliver real-time cyber threats insights (Recorded Future, 2019). The AI teams in the company cover a range of roles such as Threat Analysts, Security Specialists etc. and the awareness and understanding of the impact of algorithmic bias in the AI systems they develop is essential. Recorded Future

has 200 employees which represents a medium-sized business and the perspective of a CTO of a highly AI-driven company is relevant for the research objective.

Christian Guttmann, VP and Global head of AI at Tieto, has over 20 years of experience within the field of AI and Machine Learning. Additionally, he is executive director of the Nordic Artificial Intelligence Institute, and he could therefore provide deep expertise within the research area. Guttmann also brings in three combined perspectives which are relevant when investigating the different aspects of AI ethics. First of all, he has a technical and scientific background since he is a professor within AI. Secondly, he brings in the business perspective since he has worked within global companies such as IBM and Tieto in several countries. Lastly, Guttmann has also been interacting with and consulted governmental institutions about AI related issues. The interview with Guttmann therefore offered a unique background combination and expertise in order to gather detailed insights within the area of developing fair AI systems and mitigating algorithmic bias. As a leader of a global IT services and software company with 15 000 employees in about 20 countries, Christian Guttmann also represents a large global company's view of this topic (Tieto, 2019).

Nasim Farahini, CTO at Qamcom Research & Technology, represents a slightly smaller organization in terms of their 130 employees. Qamcom is a Centre of excellence for R&D partnerships that performs research within a range of areas, were AI related to computer vision and image recognition is a major focus area were algorithmic bias plays a vital role (Qamcom, 2019). Farahini has previously worked with computer science and development of algorithms at Google and has developed deep insights within development of algorithms and AI systems which provided value to the study in terms of her extensive expertise within the area of algorithmic bias. The final selected practitioner, a Software Engineer within Credit risk assessment (anonymous) at a Swedish IT-company, offered valuable insights into a business area which is highly vulnerable and discussed in media for introduction of unfair algorithmic bias. Since Credit risk assessments can have major implications on people's lives it is essential that the decision-making processes are transparent so that the customer can understand what factors that were considered and why a specific decision was made.

2.5 Thematic analysis

Thematic analysis is a commonly used method when conducting qualitative research, with the purpose of identifying and analysing certain patterns in collected data. Thematic analysis was used in this study as the analysis method of the data collected from the qualitative interviews since it provides a methodological way to organize and describe the data in rich detail. Thematic analysis is also beneficial for this study since it provides an opportunity to collect insights into different aspects of a research topic, which was important to fulfil the objective of the study, since the research area still is at a relatively early stage. Thematic analysis also provides an approach where the focus is

on searching for patterns across an entire data set, instead of focusing on data from a particular item or interview. Therefore, the method is suitable for this study since the data set consists of a collection of interviews that were compared against each other to find insights about similarities and differences across the data set (Braun and Clarke, 2006).

Thematic analysis can be conducted in slightly varied ways, however, in this thesis a framework with six different stages that Braun and Clarke (2006) suggests has been followed. The aim with following the steps in the framework is to form a number of relevant themes that can be used to present and group the data in the results part of the study. First of all, the verbal data in form of the recorded interviews were listen through and transcribed. The transcriptions were then read through with the aim of systematically coding the interviews for interesting and repetitive features in relation to the research questions. The codes and text extracts with similar features from each interview were then grouped together in order to gather all the data relevant to a potential theme together. When the coding process was finalized, each interview was read through again with the purpose of matching the potential themes that was found initially to the coded extracts in each interview. After this process, the seven most relevant themes in relation to the research questions was formed. The themes were then named and defined, see table 2.

After the themes were named and defined, each theme was given a colour code and the entire data collection was colour coded in relation to the themes. The extracts that were colour coded with the same theme were then grouped together and compared against the matching theme from the rest of the data set in the search for differences and similarities across the data set. When all the extracts were grouped together, a final analysis of the extracts was done, and the most relevant extracts were related back to the research questions and literature in order to present citations and to structure the presentation of the data in the results part. The themes were applied throughout the results section and the shortenings initial of each theme such as *(AU)* was used to define and mark the arguments that the respondents of the qualitative and quantitative interviews made.

Table 2: Summary and definition of the themes used in the thematic analysis

Theme	Definition
1. Awareness & understanding *(AU)*	The awareness and understanding of the challenges and potential risks with introduction of algorithmic bias in AI systems that the interviewee highlights. E.g. *"I am aware that there is discrimination against black people in the U.S. criminal justice system."*
2. Political correctness *(PC)*	The extent to which the interviewee uses political correctness and avoid expressions that could be discriminating against a group of people who are disadvantaged. E.g. *"Diversity and inclusion in our development teams is our highest priority."*
3. Authority *(A)*	The extent to which the interviewee avoids making his/her own decision about ethical dilemmas and refer to other authorities. E.g. *"It's the government or other authorities' responsibility to set up ethical regulations in order for us to follow them."*
4. Moral judgement *(MJ)*	The extent to which the interviewee makes decisions and uses arguments based on moral judgement. E.g. *"I would not treat the man stricter in this case since I think that is wrong and an unconscious bias."*
5. Technical & non-technical solutions *(TS) & (NTS)*	The technical and non-technical solutions for mitigation of bias that the interviewee mentions. E.g. *"We apply techniques for data augmentation in order to change the features artificially to represent underrepresented groups in data sets."*

6. Attitude towards own work *(AW)*	The attitude towards the interviewee's organization's work with algorithmic fairness and mitigation of bias and how it's presented. E.g. *"Honestly, we are not thinking about bias in the development process at this stage."*
7. Legal aspects *(LA)*	The legal aspects or reference to legal authorities that the interviewee mentions. E.g. *"Using the information provided to profile against gender is illegal."*

3. Results

In the following section, the results of the study are presented. The results are presented with the four different aspects covered by the research questions; The awareness and understanding of bias, the technical and non-technical solutions for mitigation of bias, the response to the scenario and finally the differences and similarities between the interviews and the online survey. The arguments that the respondents use are marked with a shortening in brackets, showing what theme from the thematic analysis that the argument is related to. This gives an overview of what kind of arguments the respondents are making and supports the discussion part of the thesis.

3.1 Awareness and understanding of bias

3.1.1 Potential risks with algorithmic bias

The awareness and understanding of potential risks with introduction of algorithmic bias in AI systems amongst the academic respondents varied. Dubhashi described that this topic is a new research area for their research group. However, he emphasized that they are aware of the problems with introduction of bias and that they investigate different methods to mitigate those, but they are by no means near a final solution. *(AW)* On the other hand, Tumova stated that the consideration of algorithmic bias in the data sets that they train their solutions on is not the core of their research at this point. At the moment they just try to make their solutions work for any data and does not prioritize balance between groups in data sets. *(AW)*

All the academics showed an understanding of that algorithmic bias in AI systems is a complex sociotechnical problem that requires both technical and social science perspectives. They also understood that the potential risks with bias are introduced since the bias in the algorithms reflect the biases that exists in our society. *(AU)* Theodorou also mentioned a technical solution regarding the importance of transparency as well as verification and validation in order to understand and identify unexpected biases. *(TS)*

> This is a complex problem that requires both engineering and socio-legal solutions. By implementing transparency, or at least through verification and validation, we can understand any unexpected biases, e.g. implicit biases, that influenced the decision making. Still, this is a short-term solution, as the real goal should always be to understand and eradicate any “bad biases” from our own culture. (Theodorou)

Both the academics and the practitioners also showed an awareness regarding the potential risks with introduction of algorithmic bias through several real-case examples of bias and unfair AI systems. Some examples that were brought up was the criminal justice system in the US, Credit risk assessments and Amazon’s recruiting algorithm, but they did not go into further details about those examples. *(AU)*

The practitioners showed a high-level of awareness and understanding and brought up risks applied on their own organization's work. Farahini, for example, showed awareness and understanding in the context of their organizations own AI systems that are used in cities to identify objects. She described that the main risks for them is misdetection of especially false negatives in the context of safety critical applications, specifically around construction areas where the consequences can be huge if the system messes up. *(AU)(AW)* Truvé shoed awareness about implicit bias since he brought up the potential risk that people are not aware that their systems are trained on unfair data. *(AU)* He also showed a deeper awareness and understanding of that there can exist both statistical bias that is necessary to separate objects in data as well as unfair and discriminating bias. *(AU)*

> There is definitely a risk with bias in that people are not aware that they are training on data which is unbalanced or not correctly collected. It can be anything from credit rating systems to the crime prediction work which the US is working on. In one way it creates unfair bias in some cases, on the other hand you can say that it's also statistically correct. (Truvé)

3.1.2 Challenges with algorithmic bias

One challenge that Dubhashi showed awareness and understanding of is that most of the bias that comes from the developer of a system are unconscious bias, which can be very difficult to identify and therefore hard to mitigate as well. *(AU)* He also described that AI systems are often seen as a black box which lacks explainability regarding why a specific decision was made. Therefore, it can be problematic for the users of the system to trust that the result is fair. *(AU)*

> The problem is probably not so much that the developers are consciously being biased, but unconscious bias will definitely seep in and where that comes from is hard to pinpoint without further study, but there will be a reflection of bias that mirrors the society. (Dubhashi)

Another challenge from the academic's point of view that both Tumova and Dubhashi highlighted is that their research groups are working in very isolated environments and that the dialogue between the academic side and different stakeholders is lacking. *(AW)* Dubhashi also presented their research groups work and described that "At the moment, we are not having dialogues with people who would use the system in a practical application or stakeholders that are in a decision-making position." *(AW)*

> The researchers tend to be very technical and like the numbers and the math's, so it's very difficult to make a bridge with social science and gender studies and the people who actually understand these issues. It's really two different roles and therefore it's very difficult to find a common speech and for us academics it's not a low hanging fruit. (Tumova)

From the practitioner's perspective, Guttmann stated that the academics are not going to completely cover the questions related to business, the governmental perspective and the impact on society. He showed awareness regarding the different resources and stakeholders that are needed in order to have an informed dialogue about this topic and to move forward within the research area. *(AU)* Truvé and Farahini also demonstrated further understanding about the challenge that the developers of AI systems face in terms of deciding on what factors they should consider in the data and what features that should be prioritized in order to develop fair systems. *(AU)*

> One of our biggest challenges is that within our entire ecosystem there is almost no one around in the Nordics were most of our customers are that do actually understand the technology and related business challenges well enough to have an informed dialogue about this topic. (Guttmann)

To summarize, the academics described their awareness and understanding of the potential risks and challenges with bias through mainly using their attitude towards their own work *(AW)* and own examples from their academic institutes. The practitioners showed a high level of awareness and understanding *(AU)* and did also use examples from their own work. However, while high-levels of awareness and understanding and attitudes towards own work *(AW)* and some technical solutions *(TS)* were described, very little moral judgement *(MJ)* and political correctness *(PC)* arguments were used, and no or little mention of legal aspects *(LA)* and authority *(A)*.

3.2 Technical and non-technical solutions for mitigation of bias

3.2.1 Data collection techniques

A majority of the respondents mentioned that the first step in order to mitigate algorithmic bias in AI systems is to make sure that their data collection methods and handling of the data sets are valid. They also highlighted the importance of their data being representative for all the groups and features within the data et in order for it to be balanced. *(AU)*

> Data collection is the most challenging part, especially for a group that is already underrepresented in open data sets. How could you for example go and find that specific minority that are underrepresented in the collected data. This is very hard and time consuming, so what we do is we try to use techniques for data augmentation. (Farahini)

Farahini described that their organization tackle the challenge with the initial data set through applying data augmentation techniques. She mentioned that "We change features artificially in our already existing data sets to be able to represent underrepresented groups and increase data sets." *(TS)* One solution they apply for achieving this is through using different photo editing tools and for example change the

colour or background of pictures and mirror images that they train their image recognition systems on. *(TS)* She makes it clear that their organization only augment the data that they have and do not apply any synthesized data methods within their data collection. *(AW)* Farahini argued that through applying these techniques within the data collection process some of the potential biases can be removed at an early stage.

3.2.2 Transparency & data visualisation tools

In order to tackle the challenge with increasing the transparency in the decision-making process of AI systems, which a majority of the respondents highlighted, several solutions were mentioned. To be able to understand how and why a particular decision by an AI system is made and when a potential bias was introduced into a development process, a level of transparency of the entire process is crucial. The anonymous respondent showed awareness regarding that decision-making processes using AI are less traceable than a traditional algorithm, which is why the need for transparency in AI systems is very high. *(AU*) The anonymous participant also highlighted that there needs to be an increased transparency considering what kind of contexts and for what purposes AI systems are used, since users often are showed very little of that. *(AU)*

> As soon as we have decision making algorithms, AI in particular, we need to create some degree of transparency so that we can understand the systems, since AI is inherently less traceable than a traditional algorithm. (Anonymous Software Engineer)

Farahini described how Qamcom actively apply data visualization tools in order for the development teams to increase the visibility and observability into their own AI models. *(AW)* Through applying these tools the teams are able to visualize the data and consider different features as well as evaluate the balance in the data sets. She also described how they apply third party technical toolkits such as the "What if" tool developed by Google and a tool called "Tcav" which increases the visibility into how different statistical features are playing a role in the decision-making process of Neural Networks. *(TS)* Their organization also continuously keep their eyes open to different features in the data that they might not think will influence the algorithm initially, but unconsciously, these features might affect the introduction of bias into the model. *(AU)* The Anonymous respondent also emphasized that they are applying technical solutions to deal with the issue "We have been testing our systems using a technical, statistical tool to check for bias. We also have discussions within our data science teams." *(TS)*

> We try to do comprehensive work for feature extraction and data visualization looking into different distributions. Based on that visualization we then get some good clues on how to compensate for the part that is not covered well. This is one way, and then we also use the data augmentation methods to compensate for the bias we find. (Farahini)

3.2.3 Validation & verification

The practitioners in particular, emphasized the technical solutions and the importance of validating and verifying their AI systems in order to maintain a fair and unbiased system over time. *(TS)* A majority of the practitioners also highlighted that fairness is an ongoing work and that performance of the model needs to be monitored continuously. *(TS)* Truvé highlighted that their organization put a lot of effort in validating their models and the historical data that they train their models on. *(AW)* They also run the system for about a month before deployment in order to inspect how the system performs and acts when new data is continuously entered into the system. *(AW)* The anonymous practitioner agreed through showing awareness and understanding about the importance of ongoing monitoring of AI systems and underlined that the perception of what is unfair might change over time and with different cultures. *(AU)*

> Ensuring fairness is an ongoing work - a system should be analysed before implementation and tested after implementation. What is unfair may also change over time and cultures. Fixing some unfairness may reveal others. Therefore, it is important for organizations to continuously work on fairness so that users can trust the systems. (Anonymous Software Engineer)

Truvé mentioned that an important part of the ongoing monitoring is the ability for the users of the systems to give feedback. *(AU)* In their AI systems they have implemented a technical solution in terms of a button in the user interface which any user can use to flag for an object that has been classified incorrectly. *(TS)* He described that "When flagging for bad classifications that data then goes back to the training team as new input data and can then be used to retrain the model." (*TS)* He also demonstrated a problem that Recorded Future call the time traveller problem which implies that if one do not understand the AI system clearly, the model can end up doing predictions about information in the past, in the sense that "The user believes that the system is doing a prediction, but what actually happened is that the user fed new real-time data into the model." *(AU)* Truvé also emphasized the importance of understanding the AI system and the data it is trained on at a deep level and to continuously verify and monitor the performance of the system to avoid this problem. *(TS)* The academics did not accentuate the importance of ongoing validation and verification of the system to the same extent as the practitioners.

3.2.4 Multidisciplinary collaboration

Regarding the non-technical aspects of solutions for mitigation of bias in AI systems, the collaboration between different disciplines ranging from engineering practice to gender studies and social science was highlighted by all the respondents. Both the practitioners and the academics agreed that collaboration across boarders is important in order to move forward within the area of mitigating algorithmic bias in AI systems. The practitioners were mainly using arguments based on awareness and understanding to

highlight this problem. Guttmann mentioned the importance of multidisciplinary collaboration in terms of initiatives were experts from different disciplines get together in order to discuss the issue. *(AU)* He described that within the Nordic Artificial Intelligence Institute the members consist of globally leading experts with a combined background in terms of technological knowledge, experience from a leading role at a large corporation and having experience from consulting the government. *(AW)* He stated that "I agree that those initiatives take place, but it is the depth of those initiatives that matters." *(AU)* The anonymous respondent agreed and underlined the significance of social scientists in the creation of AI systems. *(NTS)* The academics highlighted the lack of multidisciplinary collaboration in their academic institutions since they are working in very isolated environments. *(AW)* Tumova also highlighted a non-technical solution stating that "The gap between the technical researcher's mind-set and the social science discipline needs to be minimized in order to understand this issue fully." *(NTS)* However, the need for legislation *(LA)* and referencing to authorities *(A)* were not mentioned.

> Although ethics in AI is a relatively new topic, ethics have been discussed for a long time by social scientists. I think they need to be part of the AI creation process. (Anonymous Software Engineer)

3.2.5 Diversity in development teams

All the respondents agreed on that having diversity in terms of gender, ethnicity and culture within the development teams of AI systems will bring a wider variety of perspectives, which most likely will lead to a more fair and inclusive system with functionalities that minimize discrimination against sensitive groups.

> Garbage data in can result in garbage data out. Creators of tech are those who set the standards. The more diversity amongst developers - in terms of gender, ethnicity, culture or others - the more perspectives will be able to be included from the start, making it easier to verify unbiased systems. (Anonymous Software Engineer)

The practitioners showed their awareness and companywide initiatives on promoting diverse development teams to a wider extent than the academics, while the academics were talking more about diversity in general terms. The practitioners also promoted their own work with diversity to a higher extent than the practitioners. *(AW)* Truvé did for example mention that "Our AI development team consists of 50/50 male vs females in terms of gender diversity". *(AW)* The anonymous participant also made it clear that "We have companywide initiatives to increase diversity in our teams". *(AW)(PC)* Theodorou used a more general argument stating that "I believe that diversity brings to the table new perspectives about "what is fair" and encourages testing with data of minorities". *(PC)* He also showed awareness and understanding of that when development teams lack diversity, the emerging behaviour and the side-effects of the

system can be misread by the developers and the system can therefore be considered suitable for deployment even when it is not. *(AU)* Overall, the attitude towards own work *(AW)* and political correctness *(PC)* were mostly used to argue about the importance of diversity in the development teams to develop fair systems, while no legal aspects or regulations *(LA)* were mentioned.

3.2.6 Frameworks and guidelines

Frameworks and guidelines for regulating and guiding organization's work with AI ethics and bias were highlighted by some of the respondents. The practitioners showed attitude towards their own work when presenting their way of guiding their teams within AI ethics. Guttmann described that Tieto probably "Is one of the most advanced companies in the Nordics that are developing a structure that is necessary to make sure that the systems and the AI systems are trustworthy." *(AW)* He also mentioned a non-technical solution in terms of that their organization have published their own companywide AI ethics guidelines as well as AI certificates. *(NTS)* He described that on a first level, the primary purpose with the guidelines is to raise awareness amongst key employees dealing with AI and Machine Learning. The organization has also introduced an AI online course globally that covers different areas within AI such as technologies, ethics and bias, which all the employees should go through. *(AU)* Through raising awareness using guidelines and certificates the organization strives to be prepared for the upcoming ethical challenges with AI including bias-related problematics in AI systems. *(AU)*

Farahini agreed that there is a need to implement these kinds of standards and that for smaller companies like her organization it would be particularly helpful. *(AW)* She also suggested a non-technical solution that included working with third party organizations in order to get recommendations about tools that can be used to identify bias, remove them and implement test methods that they can use to compare their output results against. *(NTS)* She also showed a deeper awareness and understanding through underlining the importance of the ability to quantify bias in order to be able to mitigate them. *(AU)*

> If a standard procedure that you can follow is introduced, for example if there are recommended tools to apply as well as measurable metrics that you can quantify, then you could even introduce certain certifications or standards saying: this model is not biased but this one is. Then we could actually quantify these biases in a measurable way and then apply techniques to remove it. I really think we need these kinds of standards. (Farahini)

One of the academics, Theodorou, also highlighted his awareness and understanding for implementation of these kind of standards as a complement to technical solutions" Engineering solutions do not substitute the need for regulations, e.g. legislation and standards. Standards can help provide context, e.g. thresholds, or even guidelines on

how to test the system for biases." *(AU)* The other academics did not mention the need for implementation of standards and guidelines. To summarize, no legal aspects *(LA)* were mentioned in this context and most of the arguments were based on awareness and understanding *(AU)* and some non-technical solutions. *(NTS)*

3.3 Response to scenario

3.3.1 Question 1-3: Gender bias

The respondent's answers to the scenario in the qualitative interviews are presented in this section. Regarding the first question, all the participants believed that the system is gender biased and a majority agreed that it is women who are unfairly treated in this case. However, according to the guidelines, the information provided in this question is not fully sufficient to determine if the system is gender biased or not at this stage. The majority of the participants calculated the 10% false positive error rate for women and 2,5% error rate for men and pointed out that the system makes more mistakes on women and therefore they are unfairly treated. *(AU)*

However, some of the academics were not as certain that it was women who are being unfairly treated, also using arguments based on awareness and understanding, Dubhashi stated initially that the system is gender biased against women, but after some reasoning he changed his mind and stated that men are unfairly treated due to the fact that more female non-employees are being let in by the system than men. *(AU)* Theodorou argued that the system is gender biased against both genders, but more directly against women since the system fails more frequently to them. *(AU)*

Regarding the follow-up question covering what further tests the team should carry out, all the respondent's arguments were based on technical solutions. The participants stated for example that the system should be tested on a more balanced data set through adding more pictures of women for the AI system to get trained on and that the test data set should be increased overall. *(TS)* After doing that the test should be run again and evaluation regarding whether the performance and fairness would be improved or not would be performed. *(TS)* However, only one participant, Farahini, followed the guidelines for the question, through mentioning that there is no way to find out the false negative rates with this information and stated the importance of calculating that rate for the actual employees in order to decide whether the system is biased or not and what gender is being unfairly treated. *(AU)*

> The system is still gender biased if you project it on the percentages, but this time the system is more vulnerable towards women intruders than men intruders so it's actually worse for female employees but it's friendlier towards female non-employees. (Tumova)

In question 2 when the false negative rate could be calculated from the numbers in the information provided as well, the respondents mainly concluded that the system is still gender biased against women. However, some discussion regarding what unfair treatment actually meant in this case was held, like for example if it is misclassification of one gender that determine if one gender is unfairly treated or some other factors. Regarding question 3, when the false negative rates were stated explicitly, all the respondents stated that the system is clearly biased against women. Their answers were mainly based on moral judgement and the anonymous respondent argued that "This is what I think makes the system gender biased against women since it is making more mistakes on employed women than on employed men." *(MJ)* A few respondents did also propose a technical solution to the problem, Dubhashi suggested that "It would be a better comparison if you had the same number of women and you achieved this kind of difference, then it would be clearer cut." *(TS)* Farahini agreed and argued that "I think at least what you could do to identify the bias you could have this balance in your test data set." *(TS)* Farahini also emphasized that both the input data and the test data is biased in this case, which makes it even harder to identify were the bias comes from. *(AU)*

> Here we have the percentages of false negatives, so it's annoying because if you're a female employee it's harder for you to get in and that is certainly something that you would not want to have in a system like that. (Farahini)

3.3.2 Question 4: Initial vs revised system

In question 4 the respondents got presented a revised system were the false positive and false negative rates were the same for men and women. They got to reason about whether they thought the new system was better than the previous one. All the respondents used arguments based on moral judgement when agreeing on that this new system seems fairer than the initial one, but also less efficient in terms of allowing permission to the right people. *(MJ)* Truvé argued that "You could say that the system is more fair but worse. I mean if you look at the entire population it's worse but it's fairer in a sense." (*MJ)* Even though the respondents stated that the system is fairer, the majority preferred the initial system or presented a solution to how they would handle the situation instead.

> There are different ways of measuring performance and you really need to decide on the different metrics, if improving one make the other one worse then it's a trade-off and someone needs to make the decision about which system metrics we should optimize on. (Truvé)

A few respondents did not explicitly take a stand in whether they preferred the initial system or the revised one, they referred to authority in the sense that they avoided deciding about the ethical dilemma themselves and presented an alternative solution instead. *(A)* The trade-off between what metrics to optimize on was brought up by Truvé and Tumova who highlighted the ethical dilemma of whether it's acceptable or

not to make the system perform worse just to make it fairer. *(MJ)* Tumova described an alternative solution of using the initial system and make it better for women instead of making it worse for men. *(TS)* Farahini also presented a technical solution in terms of "I don't think this is a good solution, I would increase or create a balance in the data set so that you avoid this problem instead. I think this is just an artificial workaround, it's not really solving the problem." *(TS)*

However, two respondents preferred the new system using arguments based on moral judgement. Dubhashi argued that "The false positives are the same in both groups and the false negatives are also the same, so somehow it seems better." *(MJ)* Theodorou agreed and stated that he also preferred the new system "Yes, it seems better since the error rate is the same for the two populations." *(PC)*

3.3.3 Question 5: Illegal entering

When presenting question 5 to the respondents, which included information about the assumption that men are 10 times as likely to illegally enter office premises and commit crimes, the responses varied. Dubhashi stated that "If you know for sure that men are more likely to enter illegally, that information should be used when building the system and men should be treated stricter than women." *(MJ)* In the previous question he did prefer the revised system which had the same false negative and false positives for men and women, but regarding this added aspect he argued that men should be treated stricter. Truvé agreed and argued that "If this information is known, then the focus should be on reducing the number of false positives for men." *(MJ)* However, in the previous question he used arguments based on authority and avoided to decide which of the systems he would prefer. *(A)* Tumova also brought up a similar trade-off between prioritizing fairness or safety in the system and that it depends on how safety critical the system is, but she handed over the authority to someone else to make the decision. *(A)*

> Again, it comes back to what you are optimizing for. You could say that the system would be biased against men but the underlying reason for it being biased against men is because you have other statistics which tells you it's more dangerous to have a higher rate of false positives for men. (Truvé)

On the other hand, Farahini disagreed and argued that men should not be treated stricter by the system, even if you knew that men are more likely to enter illegally and commit crimes. Instead, she proposed an alternative solution of "Increase the number of images in the data set that the system is trained on and use a more balanced data set instead." *(TS)* She also mentioned the opportunities regarding that if you have a limited data set and is bound to only use 500 images as opposed to if the data set is unbounded so that you can increase the data set for both genders. *(TS)* Theodorou agreed that it is wrong to use the information when developing the system and stated that acting based on stereotypes would only introduce explicit biases in to the system, instead he proposed that "If there is a report stating that there is a higher crime rate in the area around the

office, you could introduce a second validation method, e.g. fingerprint, but for both men and women." *(TS)* The anonymous respondent also questioned whether it would be a fair system if it focused more on identifying men. *(MJ)* However, none of the respondents mentioned any legal aspects or referred to legal authorities around creating a system that has worse error rates for men on purpose, in order to make it fairer.

> I would not treat the men stricter because it's all about probabilities, I would try again to fundamentally make the system more robust and it's not by reducing the number of images for women but rather increasing the number of images for both men and women. (Farahini)

3.4 Response to online survey

The survey was answered by 303 respondents, where 63% where male, 35% female and 2% preferred not to say, see figure 7. These proportions represent 191 males and 106 female respondents. The age distribution ranged from 15 to 65+ years and the majority of the respondents belong to the age range between 25 and 54 years, with the highest percentage of 25-34 years old respondents. Regarding the educational level, a majority of the respondents hold an undergraduate degree, 38,9% followed by the second largest proportion of respondents, 34,3% which hold a master's degree. The professions of the respondents were widely spread and included both academics and practitioners. The professions ranged between student, math professor, researcher to data analyst, IT engineer etc. and the majority of the respondents have a technical role. The background information of the respondents was collected with the main purpose of viewing the distribution of the respondent's gender, age and educational level, since a skewed distribution can affect the results. In this context, the gender distribution in particular was of high importance, since the scenario covers decision making questions based on gender bias.

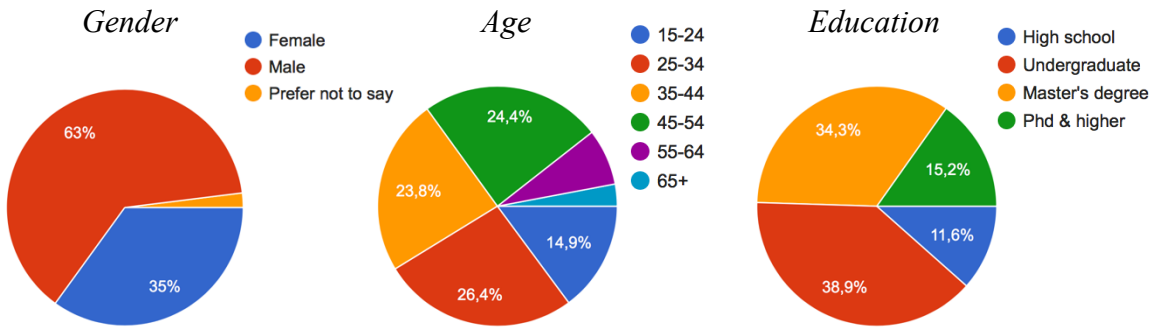

Figure 7. The gender, age and educational distribution of the respondents in the survey. Source: Own survey.

The first question in the online survey covered information about the results from the test data set and the number of non-employees that the AI system mistakenly permits entry to. Regarding the responses to question 1, the majority (63,4%) of the respondents agreed that the system is gender biased. A minority (24,1%) of the respondents did not believe that the system is gender biased and 12,5% where not sure, see figure 8. Out of the respondents that answered yes to the previous question, 78,7% believed that women were unfairly treated in this case. The respondents also got an opportunity to explain why they made their decision and a majority of the respondents mentioned that "The error rate is higher for women." *(AU)* and that "Women are unfairly treated since the system makes more mistakes on them." *(MJ)* Some respondents did also calculate the false positive rates and argued that "The 10% false positive rate for women is higher than the 2,5% rate for men." *(AU)* The minority (21,3%) of the respondents who answered that men are unfairly treated in the follow-up question mentioned arguments like, or similar to "Women gain an advantage, since they are more likely to be allowed through door, albeit wrongly." and "I don't think it's "unfair" in this case since it shouldn't let anyone in, but if I have to choose, then men have less opportunity to sneak into the building." *(MJ)* The respondents who did not believe that the system is gender biased argued that there is no unfair treatment based on gender, since this test only shows the numbers for non-employees and the actual employees are not being discriminated against. *(MJ)*

Considering the follow up question of what further tests the respondents would perform, the most common suggestion was to "Run the test again but with equal number of men and women represented in the test data set." *(TS)* and to "Use a larger sample data set to train the system on." *(TS)* Many of the respondents would also perform a test to see how the system performs for the actual employees and calculate the false negative rates. *(TS)* Another suggestion that was mentioned by a couple of respondents was to test the system for other features such as hair colour, make up etc. to see what features the system had a hard time recognizing. *(TS)*

Question 1

From the test data set of non-employees, the AI system mistakenly opens the door for 10 out of 100 female non-employees and 10 out of 400 male non-employees.

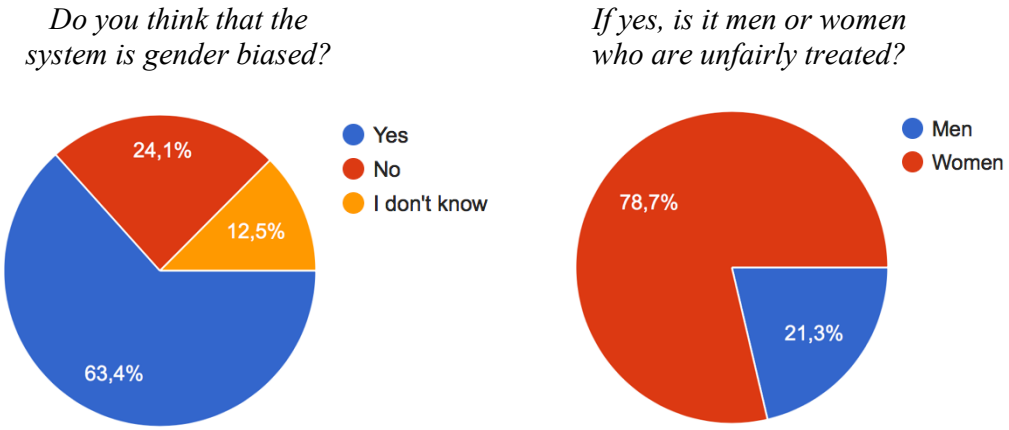

Figure 8. Responses to question 1, regarding whether the respondents believe that the system is gender biased or not and what gender that is unfairly treated. Source: Own survey.

Considering the gender distribution of the respondents in question 1 regarding whether the system was gender biased or not, see table 3A, there was a slight difference between men and women, which shows that there is a weak tendency that women believe slightly more than men that the system is gender biased. There was a similar or slightly weaker difference in question 2 and 3 (Results not shown). In the follow-up question, regarding what gender that is unfairly treated, there was no significant difference between how men and women replied to the question, see table 3B.

Table 3: Gender distribution of the respondents to question 1.

Do you think that the system is gender biased?

Gender/ Respond	**Female**	**Male**
Yes	72,6%	59,7%
No	16,0%	23,3%
I don't know	11,3%	12%

A

If yes, is it men or women who are unfairly treated?

Gender/ Respond	**Female**	**Male**
Men unfairly treated	18,5%	23,6%
Women unfairly treated	81,5%	76,4%

B

When presented question 2, considering the test where the actual employees where included, less respondents (57,1%) compared to the last question, believed that the

system was gender biased, see figure 9. There was also a higher proportion of insecure respondents (16,8%) who did not know whether the system was gender biased or not in this question compared to the previous one. However, out of the respondents who believed that the system is gender biased, a higher percentage than in the previous question (87,6%) thought that is was women who were unfairly treated in this case. Only a few respondents calculated the false positive and false negative rates mathematically in this question. The argument that a majority of the respondents had for stating that it is women who are unfairly treated was that "Women are proportionally more likely to get excluded from their workplace, since the false negative rate is higher for women." *(MJ)*

Question 2

Further tests show that the AI system permits entry to 105 women, 95 of whom are employees and 10 whom are not employees. It also permits entry to 400 men, 390 of whom are employees and 10 of whom are not employees.

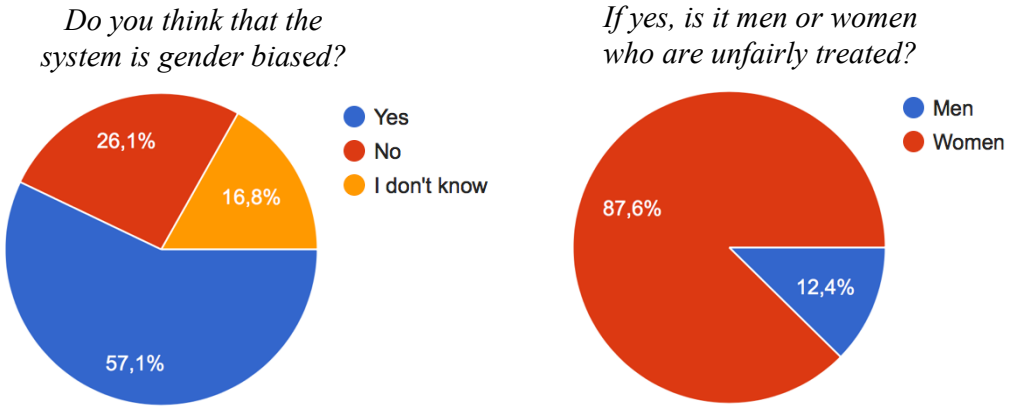

Figure 9. Responses to question 2, regarding whether the system is gender biased or not and what gender that is unfairly treated when presented additional information.

Regarding Question 3, where the false negative rates for each gender was written explicitly a majority (68,3%) of the respondents believed that the system is gender biased, which also is the highest percentage compared to the previous questions, see figure 10. Out of these respondents only 9 respondents believed that the system is gender biased against men which is equal to 4,4%, while 95,6% believed that it is women who are unfairly treated in this case. The majority of the respondents referenced to their previous answer and argued that the reason they believe that women are unfairly treated is since "Women are less likely to access their workplace successfully, since the system has a higher false negative error rate for women." *(AU)*

Question 3

Your technical team reminds you that out of the 100 female employees, 5 are not permitted entry by the system (5% false negative rate). Of the 400 male employees, 10 are not permitted entry by the system (2.5% false negative rate).

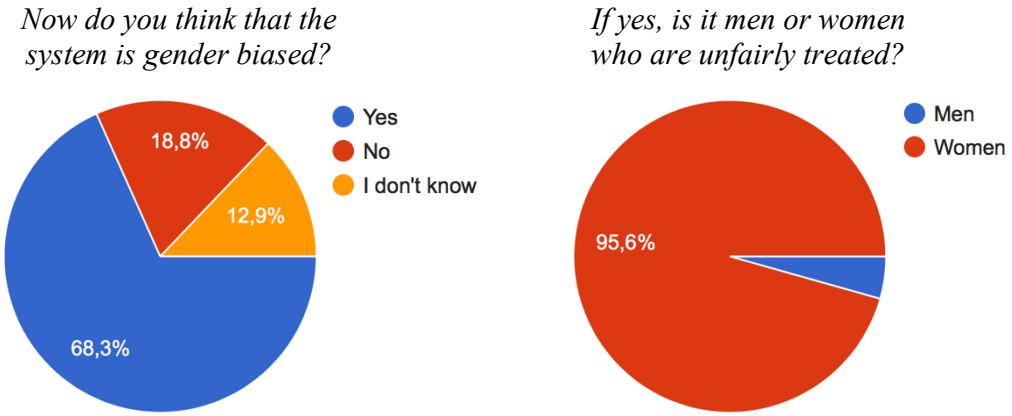

Figure 10. Responses to question 3, the respondents were presented further information regarding the false negative rates for both men and women.

The fourth question covered the decision of whether the respondents preferred the initial system or a new revised system with the same error rates for men and women. A majority (70,6%) of the participants preferred the initial system over the new one, see figure 11. Most of the respondents who preferred the initial system recognized that the revised system had a higher error rate overall and that the revised system seemed fairer, but less efficient. One respondent argued that "More errors are present in the system overall. Errors should be eliminated, not forced to be equal across gender lines." *(MJ)* Another respondent agreed that "The initial system is better in the sense that it produces fewer errors overall. The new system is less gender biased, but less good at its function." *(AU)*

A couple of respondents did also recognize that the system only performs worse for men in the revised system and that the women's error rate is the same and they suggested that the error rate for women should be improved instead of making it worse for men. *(TS)* One respondent agreed that "Degrading system functionality with respect to men hides the fundamental issue of system design for correctly identifying women." *(MJ)* Another argued that "Worse error rates so worse system. Unbiased is good, but biased and doesn't let the wrong people in is better than unbiased but insecure." *(MJ)* However there was also a proportion of respondents (29,4%) of the participants who believed that the new system seemed better. Their arguments where mostly based on moral judgement and they stated that the new system seems less biased and fairer. *(MJ)*

Question 4

After making some revisions your team announces it has improved the system so that the errors are the same for men and women. Now the system incorrectly opens the door for 10 out of 100 women and 40 out of 400 men (10% false positive error in both cases). It correctly permits entry to 95 out of 100 women and 380 out of 400 men (5% false negative error in both cases).

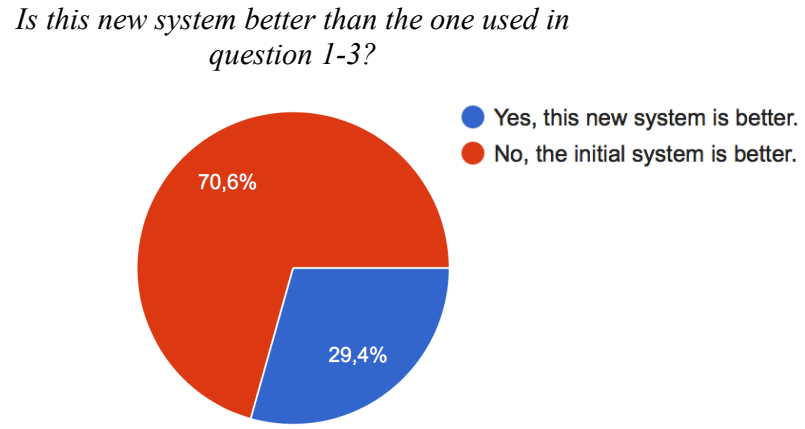

Figure 11. Responses to question 4, covering whether the respondents prefer the initial system or the new revised system.

The last question in the survey covered the aspect of deciding whether the participants would use the information provided which stated that men are 10 times as likely to illegally enter office premises and commit crimes or not. A small majority of 52,1% would use the information in the context of designing the AI system and 47,9% would not use the information, see figure 12. The majority of the respondents who would use the information used moral judgement to argue that "This information is useful, and it should be used in order to get the lowest possible error rate for male entrants in order to prevent crimes." *(MJ)* One respondent phrased the argument that "It is relevant information and means denying access to male non-employees should be minimized." *(MJ)* Another respondent agreed "Must reduce likelihood of crimes, even if using politically incorrect logic." *(MJ)* The trade-off between prioritizing safety and fairness was also brought up by one respondent stating that "Safety takes priority over fairness." *(MJ)* One respondent presented a technical solution and would use the information but focus on improving the overall performance of the system, not only for men "It's valid data used to increase the overall effectiveness of the system." *(TS)*

Out of the 47,9% who replied that they would not consider this information when designing the system, several respondents argued that it would be "illegal and unethical to do this." *(LA)* Some of them phrased that race/gender/disability profiling is illegal. *(LA)* One respondent argued that "This would introduce an unfair bias based on transient data." *(AU)* Another respondent added a new perspective considering intentional gender bias versus unintentional bias "This would be "intentional" gender

bias, and "reasoned" gender bias. What worries me is the unintentional stuff where algorithms enforce the real archaic misogyny etc." *(AU)*

Question 5
Now assume that men are 10 times as likely to illegally enter office premises and commit crimes.

Would you use this information when designing the AI system?

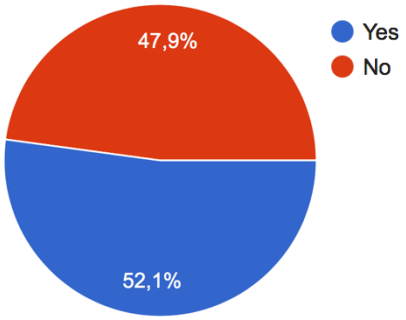

Figure 12. Responses to question 5, regarding whether the respondents would use the additional information about illegal entering or not when designing the system.

3.5 Differences and similarities in interviews and online survey

Considering the differences between the responses in the interviews and the online, if looking at the first question in the online survey a majority (78,7%) of the participants believed that the system is gender biased against women, in the second question the corresponding number was 87,6% and in third question received the highest majority of 95,6%. The number of respondents who believed the system is unfairly treating women did therefore increase depending on what information that was provided and how the information was presented. Similarly, in the qualitative interviews, the respondents got more and more certain that the system is gender biased against women.

In the interviews there were two main suggestions of further tests in question 1, increasing the data set overall and creating a balance between the genders in the training data. However, in the online survey several respondents mentioned, as in the guidelines for the scenario, that the false negative rate for the actual employees should be calculated which is essential in order to understand how the system treats the actual employees. There was only one particiapnt in the qualitative interviews who mentioned the need to calculate the false negative rate as a further test. The respondents in the online survey also proposed more varied further tests such as testing what feature the system reacted on such as testing for racial bias, make-up, glasses etc. The respondents in the online survey also based their arguments on moral judgement to a higher extent than the interviewees in question 1-3. Another difference was that in the online survey

some respondents did not think the system was gender biased at all, but in the interviews all the participants believed that the system was gender biased. The arguments from the online survey that the system is not treating any gender unfairly in question 1 was based on that this test only shows the numbers for non-employees and the actual employees are not being discriminated against, which is in line with the scenario guidelines, but it was not covered by the participants in the qualitative interviews. Regarding question 2, the main argument from the online survey responses was that the system is gender biased because women are proportionally more likely to get excluded from their workplace, since the false negative rate is higher for women. In question 3 the highest majority agreed that the system is gender biased against women, since the false negative rates are clearly stated.

Regarding the fourth question, a majority (70,6%) of the respondents in the online survey preferred the initial system, similarly in the qualitative interviews, a majority also preferred the initial system. The interviewees mainly based their arguments on moral judgement and provided technical solutions to the problem in terms of using the initial system but make it more gender balanced so that the accuracy can be improved for women instead of making it worse for men. Some interviewees also avoided to decide what system they preferred, using authority related arguments, and instead highlighting the trade-off between fairness and accuracy instead of deciding themselves. In the online survey, there were no arguments based on authority, they were similarly based on moral judgement and technical solutions to the challenge.

Regarding the fifth question in the online survey, there was a small majority of 52,1% who would use the information in the context of designing the AI system and 47,9% who would not use the information. The respondents that would use the information based their arguments mainly on moral judgement and highlighted that the likelihood of crimes must be reduced, even if using politically incorrect logic. The arguments that the respondents who would not use this information used was based on moral judgment and several respondents also mentioned the legal aspects, stating that it is illegal with profiling based on gender. However, in the qualitative interviews the respondents did not mention the legal aspects, they used a reasoning based on moral judgment and technical solutions instead. In the qualitative interviews, only a few persons stated that they would use the information to treat men stricter. The other participants either used arguments related to authority and highlighted the dilemma of that the decision depends what you are optimizing for, fairness or safety or provided alternative technical solutions to the problem.

4. Discussion

In the following section, discussion of the findings from the results section is presented. The discussion follows a similar structure as the result, with one section covering topics related to each research question. The same notations and references to the thematic analysis as in the results section is applied in this section.

4.1 Awareness and understanding of bias

All the practitioners showed a high level of awareness and understanding *(AU)* of potential risks and challenges with bias, while the academics used more attitude towards their own work *(AW)* and low-level examples of algorithmic bias in society, without showing a deeper level of understanding. Despite a high level of awareness and understanding *(AU)* and attitude towards own work *(AW)* was shown, neither the practitioners or the academics mentioned or talked very little about legal aspects *(LA)* and they referred very little to authorities *(A)* when they presented their own work.

Overall, the practitioners showed a higher level of understanding and awareness than the academics and they had even started to implement and use tools to mitigate algorithmic bias, while the academics stated that they are at an early state regarding this area. This indicates that there is a need for the academic side to collaborate more extensively with practitioners. Tieto has for example implemented their own AI ethics guidelines, which shows effort in order to create awareness amongst their employees about the issue of bias and ethical dilemmas in AI systems (Tieto, 2018) which is what smaller organizations and academics are asking for. Regarding the potential risks with algorithmic bias in AI systems, the academics and the practitioners showed a high level of awareness about both real-world examples and an understanding for that the problem is complex and that it is a sociotechnical problem which needs to be dealt with through collaboration with stakeholders and across boundaries. The main challenges that were brought up from the academics were that they function in an isolated area and do not have much external contact with stakeholders or industry representatives that has a decision-making power regarding this area. From a global organization's point of view their main challenge is that there is almost no one around in their ecosystem within the Nordics that know this area well enough to have an informed dialogue on the topic, which indicates the importance for multidisciplinary collaboration in order to move forward within the area.

4.2 Technical and non-technical solutions for mitigation of bias

When presenting solutions for mitigation of bias, the practitioners showed a lot of awareness and understanding *(AU)* and used arguments related to their attitude towards own work *(AW)* and they also provided more details about technical solutions *(TS)* that

they apply in order to mitigate bias, than the academics. Yet, very little political correctness *(PC)* and moral judgement *(MJ)* and almost no legal aspects *(LA)* were mentioned. The technical solutions that were brought up for mitigation of algorithmic bias in AI systems were mainly covering three areas; *Data collection methods*, *Transparency & data visualisation tools* and *Validation & verification*. The practitioners clearly stated their own organization's work and described that they are implementing tools for mitigation of bias in practice; such as using data augmentation methods in order to increase data for underrepresented groups or testing their systems for bias through validation and verification for at least a month before deployment to ensure that the system is trained on inclusive data and that it gives accurate results. Regarding the non-technical solutions, the areas that were covered were; *Multidisciplinary collaboration*, *Diversity in development teams* and *Frameworks & guidelines*. Again, the practitioners used arguments based on their attitude towards their own work when describing that multidisciplinary collaboration is critical to move forward within this area and some of the organizations stated that they have implemented AI ethics guidelines and are working actively to raise awareness about this issue amongst their employees. However, the academics did only mention that they are investigating methods to deal with this issue, but they are not at an implementation state yet. This shows that practitioners are several steps ahead than academics when it comes to actual implementation of both technical and non-technical solutions for mitigation of bias in AI systems.

However, all the respondents mentioned different technical tools or non-technical methods as their own way to deal with this issue, which is similar to the idea that IBM developed their own platform, IBM AI OpenScale, to handle mitigation of bias in AI and ensure development of fair AI systems (Howard and Howard, 2018). This indicates that at this stage, many organizations that are active within AI, develop their own tools and frameworks, since there are no clear general guidelines or frameworks that are implemented across organizations yet. This implies that there is clearly a need for stakeholders to work together across boundaries in order to provide some general guidance within this area.

4.3 Response to scenario

Considering how the participants responded to a practical scenario presented to them, the study shows that when the interviewees where answering the questionnaire they showed high levels of awareness and understanding *(AU)* of the problem and used their attitude towards their own work *(AW)* when presenting their organizations. The academics showed slightly less that they had put technical solutions into practice and showed more of their academic work. However, when talking in general about their work neither the practitioners or the academics used much arguments based on moral judgement *(MJ)* and they did not use political correctness. *(PC)* They also mentioned no or very little legal aspects. *(LA)* However, when presented a practical scenario, the

participants very quickly went over to use moral judgement *(MJ)* to solve the questions, they continued to use arguments based on awareness and understanding *(AU)* when responding to these questions and several participants did also propose technical solutions *(TS)* such as increasing the test data set and creating more balance between the genders in the training data. Again, they referred very little to legal aspects and authorities. *(LA)(A)*

The fact that no legal aspects were discussed, implies that the organization's focus at the moment is not legal aspects regarding this issue, they focus more on developing and using their own models and frameworks than on regularization in order to handle the problem. This goes in line with the theory about disruptive technology presented in the introduction (Kolacz and Quintavalla, 2019). This also indicate that AI can be viewed as a disruptive technology, and that the legal aspects often occur after widely use and adoption of the technology. Therefore, there is also a risk that AI, in the context of a disruptive technology, can ignore the legal frameworks and organizations can identify loopholes in the law which makes it possible for the technology to get distributed quickly and spread amongst several markets before the law and legal authorities are catching up, like with the Uber example (Isaac and Davis, 2014).

4.4 Differences and similarities in interviews and online survey

Considering the differences and similarities in how people in different organizations and individuals responded to an online survey, a majority agreed that the system is gender biased in question 1-3 of the scenario and that women are unfairly treated. In the qualitative interviews both the academics and the practitioners used their awareness and understanding *(AU)* of the topic to argue why the system is gender biased and against what gender. A couple of respondents did also propose technical solutions *(TS)* as arguments for their reasoning. In both the interviews and online survey, the respondents increased their confidence in that women are unfairly treated depending on the information that was provided. Regarding question 4, the majority of the respondents in both the qualitative and quantitative part preferred the initial system with higher accuracy even though they stated that the new system seemed fairer and less gender biased.

The differences in responses regarding the fifth question showed that the participants in the qualitative studies did not base their arguments on legal aspects *(LA)*, while several respondents in the online survey mentioned that it might be illegal to use profiling based on gender to treat men stricter. This shows that the question was asked clearly enough for it to be relevant to consider the legal aspects, but the participants in the qualitative interviews did not. This also indicates that the participants in the qualitative interviews based their arguments on awareness and understanding *(AU)* and moral judgment *(MJ)* and did not think about legal aspects *(LA)* in the used scenario. Both the practitioners

and the academics showed their awareness and understanding of the topic clearly in the interviews, but then again, they did not mention any legal aspects. The conclusion based on that information is therefore that when it comes to mitigation of bias in AI systems, the legal aspects can get hidden behind moral judgment and people's own awareness and understanding of the area. Hence, a risk for AI systems in the context of a disruptive technology, is that legal aspects and regulations are ignored. This is an insight which is important to consider when developing AI systems in order to make sure that the systems that are developed are fair and inclusive. This results also implies the need for establishing new AI ethics roles along with development of fair AI systems, such as transparency engineers and AI trainers (Tieto, 2018). It also highlights the need for an AI ethics role who can ensure that legal aspects do not get ignore and that they are considered when developing AI systems in order to make the systems fair and inclusive.

5. Conclusions

The objective with this study is to provide insights regarding different aspects that are important to consider in order to mitigate algorithmic bias as well as to investigate the practical implications of bias in AI systems. To fulfil the first part of the objective a normative approach was used in order to describe how fair decisions are ought to be made. The second part of the objective was fulfilled through applying a descriptive approach to algorithmic fairness where a practical scenario was applied, and respondents were asked questions regarding their own understanding of gender bias in a face recognition scenario. The first two research questions were covered through a questionnaire and the following two were covered using a practical scenario.

The first research question covered to what extent organizations are aware of and understand the potential risks and challenges with algorithmic bias in AI systems. The results showed that the practitioners showed a higher level of awareness and a deeper level of understanding than the academics. The academics used a lot of examples from their own work and academic institutions, while the practitioners used their awareness and understanding to describe their work and awareness in more detail. However, even though the interviewees showed a high level of awareness and understanding and attitude towards own work, there was very little awareness showed regarding legal aspects and no or little references to legal authorities.

The second research question covered the technical and non-technical solutions that can be implemented to mitigate algorithmic bias in AI systems. The main methods that were brought up was; *Data collection methods*, *Transparency & data visualisation tools* and *Validation & verification* as technical methods and the non-technical methods were; *Multidisciplinary collaboration*, *Diversity in development teams* and *Frameworks & guidelines*. The study also shows that practitioners are several steps ahead of the academics when it comes to actual implementation of both technical and non-technical solution for mitigation of bias in AI systems and that organizations tend to develop their own solutions and frameworks rather than adopting to general standards.

The third research question covered the practical implications of algorithmic bias in AI systems and how people in different organizations respond to a practical scenario related to algorithmic bias presented to them. The study shows that when the interviewees presented their work in general terms, they mostly used their awareness and understanding and their attitude towards own work as arguments. However, when presented a practical scenario, they quickly switched and used mainly moral judgment to solve the scenario. They continued to use examples from their own work and presented technical solutions to the problem, but they referred very little to legal aspects and authorities. This concludes that AI can be seen as a disruptive technology, since organizations tend to develop their own solutions without being aware or certain about the legal consequences and therefore legal aspects tend to fall behind. There is therefore

a risk that legal aspects get hidden behind organizations own moral judgments and technical solutions for mitigation of algorithmic bias in AI systems.

The final research question brought up the differences and similarities between the interviews of people with various roles in AI and an online survey covering a practical scenario. The main conclusion regarding this question is that from the online survey, where a wider range of potential answers to the practical scenario was shown, it was clear that there was no reason that legal aspects should not be included in the responses, since several respondents in the online survey raised the legal aspects when answering this question. Hence, this also implies the risk for AI systems in the context of a disruptive technology, that legal aspects and regulations are ignored, which is important to consider when developing AI systems in order to develop fair AI systems.

5.1 Future work

In this study, seven qualitative interviews were done with both practitioners and academics active within the field of AI. Based on the results of the study, it would be interesting to extend the qualitative interviews and interview more persons that are specifically involved within the legal aspects and frameworks of mitigation of bias in AI systems. In that way, further conclusions regarding the risk of seeing AI in the context of a disruptive technology could be drawn and an investigation of what legal regulations that are being developed in parallel with the increased use of AI systems in society could be done. Further work considering the need for new AI ethics roles within development teams of AI systems could also be studied, such as the need for and the practical implications of AI ethics roles with responsibility over the legal aspects when implementing new AI systems.

Regarding the practical implications of bias in AI systems, the scenario that was used in this study covered the area of face recognition and gender bias. As a suggested future work, it would be interesting to investigate whether a change of the context of the scenario to another type of AI system such as a credit risk assessment system or a criminal justice system would result in similar or different answers regarding the fairness of the system and who is unfairly treated. Another suggestion regarding further work considering the practical scenario could be to use another sensitive feature than gender, such as ethnical background or skin colour, but within the same context and similar questions as in this study. Then evaluation of whether the responses would differ or remain similar compared to the responses to the scenario in this study could be done. A few different scenarios could for example be used with various contexts and different sensitive features and the responses to questions regarding these systems could then be compared against each other in terms of similarities and differences.

References

Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016), ”Machine Bias – There’s software used across the country to predict future criminals. And it’s biased against blacks”, ProPublica. Available online: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (2019-03-04).

Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards, J.T., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., & Zhang, Y. (2018), AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. CoRR.

Braun, V., & Clarke, V. (2006), “Using thematic analysis in psychology”, Qualitative Research in Psychology, Vol. 3, No. 2, 77-101.

Brownstein, M. (2016), Implicit bias. In E. N. Zalta (Ed.), The Stanford encyclopedia of philosophy. Available online: https://plato.stanford.edu/archives/win2016/entries/implicit-bias/ (2019-04-20).

Chalmers (2019), Chalmers AI Research Centre, About us. Available online: https://www.chalmers.se/en/centres/chair/About-us/Pages/default.aspx (2019-03-10).

Christensen, C. M. (1997), The innovator's dilemma: When new technologies cause great firms to fail. Boston, Mass: Harvard Business School Press: Boston.

Danks, D., & London, A.J. (2017), Algorithmic Bias in Autonomous Systems. In Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017) forthcoming.

Dastin, J. (2018), “Amazon scraps secret AI recruiting tool that showed bias against women”, Reuters. Available online: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08GG (2019-05-20).

Dieterich, W., Mendoza, C. & Brennan, T. (2016), COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Technical Report, Northpointe Inc.

Dignum, V. (2017), Responsible autonomy. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI’2017), pp. 4698–4704.

Dressel, J. & Farid, H. (2018), "The accuracy, fairness, and limits of predicting recidivism", Science advances, Vol. 4, No. 1, pp. eaao5580.

Garcia, M. (2016), "Racist in the Machine: The Disturbing Implications of Algorithmic Bias", World Policy Journal, Vol. 33, No. 4, pp. 111-117.

Graziano, A. M. & Raulin, M. L. (2013), Research methods: A process of inquiry. Vol. 8. Pearson: Boston.

Grgic-Hlaca, N., Redmiles, E.M., Gummadi, K.P., & Weller, A. (2018), Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. WWW.

Howard, A. & Borenstein, J. (2018), "The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity", Science and engineering ethics, Vol. 24, No. 5, pp. 1521-1536.

Howard, P. & Howard, D. (2018), IBM AI OpenScale. Bloor Research. Available online: https://www.ibm.com/downloads/cas/RW0GR5D77 (2019-03-15).

IBM Policy (2018), “Bias in AI: How we Build Fair AI Systems and Less-Biased Humans”, Available online: https://www.ibm.com/blogs/policy/bias-in-ai/ (2019-02-20).

Isaac, E., & Davis, U. C. (2014), “Disruptive innovation: Risk-shifting and precarity in the age of Uber”, Berkeley Roundtable on the International Economy, University of California, Berkeley.

Kamishima, T., Akaho, S. & Sakuma, J. (2011), "Fairness-aware Learning through Regularization Approach", IEEE, pp. 643.

Kolacz, M.K. & Quintavalla, A. (2019), "Law in the Face of Disruptive Technology, An Introduction", European Journal of Risk Regulation, Vol. 10, No. 1, pp. 1-3.

Larson, J., Mattu, S., Kirchner, L. & Angwin, J. (2016),” How we analysed the COMPAS Recidivism Algorithm”, ProPublica. Available online: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (2019-04-20).

Merler, M., Ratha, N., Feris, R. & Smith., J.R. (2019), Diversity in Faces. IBM Research AI.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014), Qualitative data analysis: A methods sourcebook. 3th Edition. Sage: Los Angeles.

Qamcom (2019), Who we are. Available online: https://www.qamcom.se (2019-03-11).

Recorded Future (2019), Intelligence-Driven. Security. Available online: https://www.recordedfuture.com (2019-03-10).

Srivastava, B. & Rossi, F. (2018), "Towards Composable Bias Rating of AI Services", pp. 284. AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 12/2018.

Tieto (2018). Tieto strengthens commitment to ethical use of AI. Available online: https://www.tieto.com/en/newsroom/all-news-and-releases/corporate-news/2018/10/tieto-strengthens-commitment-to-ethical-use-of-ai/ (2019-05-15)

Tieto (2019). About us, our company. Available online: https://www.tieto.com/en/about-us/our-company/ (2019-03-12).

Interviews

Anonymous; Software Engineer – Credit risk assessment. IT Company. (2019), Written interview, Stockholm, 19 April.

Dubhashi, Devdatt; Professor, Data Science & AI. Chalmers University. (2019), Personal interview, Lindholmen Conference Centre, Gothenburg, 5 March.

Farahini, Nasim; CTO – AI, IoT & Cloud. Qamcom. (2019), Personal interview, Kista, 4 April.

Guttmann, Christian; VP & Global Head of AI, Tieto. (2019), Phone interview, Kista, 27 March.

Theodorou, Andreas; Researcher, Artificial Intelligence - Responsible AI. Umeå University. (2019), Written interview, Stockholm, 29 April.

Truvé, Staffan; CTO & Co-founder. Recorded Future. (2019), Video interview, Uppsala, 18 March.

Tumova, Jana; Professor, Robotics. Royal Institute of technology. (2019), Personal interview, Royal Institute of Technology, Stockholm, 26 March.

Appendix A

Qualitative questionnaire

- What type of bias or fairness challenges related to AI do you see as most relevant in your specific business/research context?
- What potential risks do you see with introduction of algorithmic bias in AI systems and to what extent are you and your team aware of these risks?
- What kind of stakeholder engagement or initiatives do you see as necessary to ensure that AI systems are developed in a fair and inclusive way?
- When an AI system has made a decision, how can you make sure that the user can trust that the result is fair?
- Do you apply any practical steps in order to mitigate unfair bias in the development process of AI systems? If yes, which methods do you use? (It can be either technical steps and/or non-technical regulations).
- How do you handle the data sets that AI systems are trained on to ensure it is used in a way that takes potential concerns about bias or inclusion in to account?
- How do you work with diversity in the development teams of AI systems and how do you think that affects how fair the final system becomes?
- Do you think that the fact that more developers are men, introduces bias into AI-systems? Why/why not?

Appendix B

Quantitative online survey

Avsnitt 1 av 7

Scenario AI system for face recognition

Dear respondent,

I'm a master student in Sociotechnical Systems Engineering at Uppsala University. I focus my Master Thesis on studying how AI systems can be developed in a fair and inclusive way as well as how unfair bias in AI systems can be mitigated. As part of the study a practical scenario about an AI system for face recognition is being used. I would really appreciate if I could borrow 5 min of your time to complete the scenario. Your answers will remain anonymous.

Thank you very much for your help!

Bildrubrik

Gender *

- ○ Female
- ○ Male
- ○ Prefer not to say

Age *

- ○ 15-24
- ○ 25-34
- ○ 35-44
- ○ 45-54
- ○ 55-64
- ○ 65+

What's your profession? *

Kort svarstext

Educational level *

- ○ High school
- ○ Undergraduate degree
- ○ Master's degree
- ○ Phd & higher

Description of scenario:

An AI system has been developed to automatically open the door of a building for employees using face recognition. The system uses a database of all 500 employees, 100 of whom are women and 400 are men, plus a test data set of 500 non-employees (also 100 women and 400 men).

The system is required to have an error rate (both false positives and false negatives) of less than or equal to 10% to reach an acceptable operational standard.

Consider the following 5 questions about the system. There are no definitive right or wrong answers, I just want to understand your reasoning about these problems.

Avsnitt 3 av 7

Question 1

From the test data set of non-employees, the AI system mistakenly opens the door for 10 out of 100 female non-employees and 10 out of 400 male non-employees.

(Definition of gender bias: Unfair difference in the way women and men are treated)

Do you think that the system is gender biased? *

- ○ Yes
- ○ No
- ○ I don't know

If yes, is it men or women who are unfairly treated?

- ○ Men
- ○ Women

Why? Please explain shortly.

Lång svarstext

What further tests would you ask your team to carry out? *

Lång svarstext

Avsnitt 4 av 7

Question 2

Further tests show that the AI system permits entry to 105 women, 95 of whom are employees and 10 who are not employees. It also permits entry to 400 men, 390 of whom are employees and 10 of whom are not employees.

Do you think that the system is gender biased? *

- ○ Yes
- ○ No
- ○ I don't know

If yes, is it men or women who are unfairly treated ?

- ○ Men
- ○ Women

Why? Please explain shortly.

Lång svarstext

Avsnitt 5 av 7

Question 3

Your technical team reminds you that out of the 100 female employees, 5 are not permitted entry by the system (5% false negative rate). Of the 400 male employees, 10 are not permitted entry by the system (2.5% false negative rate).

Now do you think that the system is gender biased? *

- ○ Yes
- ○ No
- ○ I don't know

If yes, is it gender biased against men or women?

- ○ Men
- ○ Women

Why? Please explain shortly.

Lång svarstext

Avsnitt 6 av 7

Question 4

After making some revisions your team announces it has improved the system so that the errors are the same for men and women. Now the system incorrectly opens the door for 10 out of 100 women and 40 out of 400 men (10% false positive error in both cases). It correctly permits entry to 95 out of 100 women and 380 out of 400 men (5% false negative error in both cases).

Is this new system better than the one used in questions 1-3? *

- ○ Yes, this new system is better.
- ○ No, the initial system is better.

Please provide a short explanation of your answer.

Lång svarstext

Avsnitt 7 av 7

Question 5

Now assume that men are 10 times as likely to illegally enter office premises and commit crimes.

Would you use this information when designing the AI system? *

- ○ Yes
- ○ No

Please provide a short explanation of your answer.

Lång svarstext