NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

# Biased Artificial Intelligence

## Algorithmic Fairness and Human Perception of Biased AI

Sidney Anna Machill

Dissertation presented as the partial requirement for obtaining a Master's degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**
Universidade Nova de Lisboa

# BIASED ARTIFICIAL INTELLIGENCE

Algorithmic Fairness and Human Perception of Biased AI

Sidney Anna Machill

Dissertation presented as the partial requirement for obtaining a Master's degree in Information Management, Specialization in Information Systems and Technologies Management

**Advisor:** Flávio Luís Portas Pinheiro & Diana Orghian

July 2020

# ACKNOWLEDGEMENTS

I would like to appreciate my deep and sincere gratitude to my supervisors Diana and Flavio as well as my family and friends.

# ABSTRACT

As Artificial Intelligence (AI) is given more power in many decisions, potential resulting biases in respect to gender, race, and other minorities have to be analyzed and reduced to a minimum. Machine Learning (ML) models are implemented in various areas and can decide who gets invited to an interview, granted a loan, gets the right cancer treatment, or goes to prison. Consequently, biases can have a crucial negative impact on people's life. This thesis highlights previous research in this field, shows its limitations and breaks down the content into its core components in a systematic manner. Therefore, types of existing biases, and areas where AI bias is most prevalent are defined. Further, root causes for discriminating algorithms are analyzed according to the AI model creation chain: data, coder, model, and model usage. An abundance of fairness measurements is classified and elaborated in a tabular format. Thereafter, bias mitigation techniques naming pre-processing, in-processing, and post-processing for ML algorithms are summarized, critically analyzed and limitations of research for unsupervised learning fairness measures are indicated. In addition, a survey is conducted analyzing how people perceive AI and human biases in various scenarios. The relation of their gender, occupation, AI/ML knowledge and experience with the risk assessment of biases is analyzed. Three hypotheses are statistically tested which analyze the difference in bias risk perception of human decision in comparison to AI models, the impact of biased AI training and team diversity on the risk awareness of biases. For various scenarios significant findings were found that show a higher risk perception of the human scenarios in comparison to AI scenarios. While no significant impact of biased AI training exists, race and disability diversity in teams have shown positive impact on risk aversion towards biases.

# KEYWORDS

Artificial Intelligence; Biased Artificial Intelligence; Machine Learning; Algorithmic Fairness; Algorithmic Perception; Discrimination

# INDEX

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| **ADA** | American with Disabilities Act |
| **ADEA** | Age Discrimination in Employment Act |
| **AFR** | Automated Facial Recognition |
| **AI** | Artificial Intelligence |
| **ATMs** | Automated Teller Machines |
| **BAME** | Black, Asian, and other minority ethics |
| **CRISP-DM** | Cross-industry standard process for data mining |
| **DVLA** | Driver and vehicle licensing authority |
| **EEOC** | Equal Employment Opportunity Commission |
| **INTERPOL** | International Criminal Police Organization |
| **BKA INPOL** | German Federal Criminal Police Office Information Systems |
| **LGBT** | Lesbian, gay, bisexual, transgender |
| **LGTQ** | Lesbian, gay, bisexual, transgender, queer |
| **m** | Mean |
| **ML** | Machine Learning |
| **ROI** | Return on Investment |
| **SAT** | Scholastic Aptitude Test |
| **sd** | Standard Deviation |
| **STEM** | Science, Technology, Engineering, Mathematics |
| **U.K.** | United Kingdom (of Great Britain and Northern Ireland) |
| **U.S.** | United States (of America) |

# 1. INTRODUCTION

The rising influence of Artificial Intelligence (AI) is unquestionable and irreversible. 43 percent of surveyed companies implemented at least one kind of AI in their enterprise and over 40 percent of questioned customers expect AI to expand access to various areas such as finance, legal, transportation, and healthcare (PwC, n.d.). Data has been described as the "new gold" and at present, AI systems can actually create value from the abundance of data we have produced and stored. AI systems can learn from past human behavior and "reason" in a logical manner. By taking in past data, models learn from past patterns and project these findings into the future. AI certainly brings many benefits, but does it also come with negative impacts for our societies?

Recent cases of discriminating AI systems have drawn public attention on the topic of biased AI. Human biases can be encoded into the systems and affect forthcoming decisions. The consequence of these biased decisions can be severe, and can, for instance, influence decisions on prison time, loan granting, or hiring. However, just 47% of companies test for bias in data, models, or the human application of AI algorithms, leaving people at risk for discrimination and unfairness. Minority groups such as women, people of color, people from the LGBTQ community or people with disabilities, among others, are especially vulnerable. The known cases are, however, just the tip of the iceberg as the systems are non-transparent and victims can be unaware of their doom.

As AI systems are constantly growing in importance and becoming embedded in the most private and important aspects of our lives, influencing millions of people (West, Whittaker & Crawford, 2019), so should the awareness of biases in AI grow. This thesis aims to start right there by giving an overview of biased AI and many of its aspects. Therefore, I will introduce the topic of AI and explain the fundamental conceptions of bias, fairness, and discrimination. Following, five types of biases are explored: race, gender, age, sexual orientation, and disability. Their modality, implications, and most common areas of existence are expounded, namely employment, finance, criminal defense, and healthcare. Further, I explore the causes why biases can exist in AI models. These causes are organized according to the development chain of an AI system: first, the data, followed by the coder, the model, and finally the model deployment. Thereafter, solutions for biased AI are elaborated including an overview of fairness definitions and metrics to detect and measure bias as well as various bias mitigation techniques. The latter part is again structured among the algorithm development chain into pre-processing, in-processing, and post-processing approaches.

Finally, a survey is carried out to assess people's perception of AI biases. Understanding those perceptions is essential to mitigation and developing precaution measures against those biases. A good benchmark to understand how people perceive algorithmic biases is by comparing them with human biases. In the present survey we compare perception of risk of bias in actions performed by humans and actions performed by AI systems. Thereby, 205 individuals are questioned to assess the risk, reason of bias, and potential mitigation techniques of various AI and human scenarios. Three hypotheses are developed analyzing risk perception of human biases and AI biases, the effectiveness of biased AI training on the risk assessment, and the influence of team diversity on the risk perception of biases.

With this thesis I hope to shed light on a complex and sensitive topic in this new technological era in which algorithms, machines, and robots are making many decisions on our behalf. Alongside an

extended literature review, I provide data about risk perceptions of biases in AI, and with that I hope to bring further awareness about the need for systematic research on the causes and mitigation strategies, but also about the need for clear guidelines for industry and companies developing this kind of technologies on how to test their systems for biases and ways to correct it.

## 2. METHODOLOGY

The research of this thesis is conducted in the form of a critical literature review which intends to be an objective, summarizing, and critical analysis of scientific research available about biases in AI (Cronin, Ryan & Coughlan, 2008; Hart, 2018). This purpose aligns with the specific objective of providing the reader the opportunity to become familiarized and reflect about the important aspects of biased AI by summarizing an abundance of literature available.

The process of a literature review follows roughly five steps including the topic selection, the search for literature, the gathering and exploring of literature followed by the writing of the review and references (Cronin et al., 2008). First, the topic selection was done to fulfill the need to have a comprehensive overview of a socially relevant problem that requires spread awareness for an effective solution to be achieved. Second, the search for literature was conducted by reviewing what I considered to be the most important studies, articles, books, and papers. Most of the literature examined for the purpose of this thesis was published within the last five years due to the recent nature of the topic at hand. However, older literature is also used as it touches on relevant concepts and insights from psychology, law, politics, or computer science. Around 450 sources were gathered and examined whereby roughly 300 sources were analyzed and summarized in writing.

Various frameworks exist aiming to define the type of literature review. One form of classification is presented by the University of Southern California (n.d., see also Baumeister & Mark, 1997) whereby they differentiate between an argumentative, integrative, historical, methodological, theoretical, and systematic review. In their understanding, this literature review can be seen as an integrative review. This type of review takes literature and integrates it in such a way that new frameworks are generated enabling new perspectives on the subject matter. It is integrative also because it takes existing research in different fields, such as psychology, law, data science, and social science and combines them, gaining a more holistic perspective of the problem of biased AI. The thesis structures the topic of biased AI in its facets, and therefore, provides a new view on the problem. Thereby, the thesis roughly divides literature into four main parts: types of biases; areas where biased AI is prevalent; reasons for AI being biased; and solutions including fairness measures and mitigation techniques. These four parts are subdivided into smaller subsections to provide a clearer structure as well as a dictionary-like purpose to focus on certain relevant topics. The first two parts are closely intercorrelated as different types of bias occur in specific areas.

The Figure 1 and Table 2 were created in line with the concept of an integrative literature review as new frameworks are developed which not only summarizes literature but also contributes to the topic with a new perspective. Finally, to better understand how people perceive AI in the context of biased decisions in different areas, a survey is conducted analyzing the risk assessment in a set of scenarios. The methodology concerning this part can be found in section 8.2.

To sum up this section, the general research questions asked and hope answering with this thesis, are the following:

> “What kind of biases in AI systems do exist?”
>
> “In which areas are AI biases most prevalent?”

“Why are AI systems biased?”

“How to measure algorithmic fairness?”

“How to mitigate biases in AI systems?”

“How do people assess the risk of biases in AI, when compared to other humans?

# 3. FUNDAMENTALS

## 3.1. ARTIFICIAL INTELLIGENCE

Already in the first half of the 20th century, intelligent robots occurred in science fiction as in the Wizard of Oz and in Metropolis. In the mid-century, scientists like the British polymath Alan Turing (1950) had a clearer vision of AI. In a paper called “Computing Machinery and Intelligence” from 1950, Turing explored the mathematical potential behind the fictional concept of a thinking machine. In his understanding, machines could imitate the human capacity to refer to available information and reason in order to make decisions. Ahead of his time, he could not prove his concept as computers were, inter alia, lacking computational power and extremely expensive. Five years later, the first AI program “Logic Theorist” was initialized and presented at the Dartmouth Summer Research Project on Artificial Intelligence. The conference ignited the research on AI in the following years (Anyoha, 2017).

From there on, AI flourished, became faster, cheaper, and more accessible for everyone (Anyoha, 2017). In 1970, Life Magazine quoted Marvin Minsky, co-founder of the Massachusetts Institute of Technology's AI laboratory, stating “from three to eight years, we will have a machine with the general intelligence of an average human being” (as cited in Time Inc., 1970, p. 58). Even though the proof of concept was achieved, much research and development had still to be done to succeed in the area of natural language processing, self-recognition, and abstract thinking (Anyoha, 2017). Even if AI research did not receive much government funding, it thrived between the 1990s and 2000. Especially in 1997, AI received much publicity when chess champion Gary Kasparov was defeated by IBM’s AI-driven chess-playing computer program Deep Blue.

Today, we live in a “big data” era. We have the computational and storage capacity to collect, process, and leverage the enormous abundance of data that humans produce. These circumstances contribute to massive growth in AI and, therefore, AI is, as Microsoft's Chief Officer Dave Coplin states, “the most important technology that anybody on the planet is working on today” (Furness, 2018, para. 1). Now, AI systems are embedded in our everyday lives and are prevalent in many areas: micro and macro financing, health care, education, crime, human resources, and marketing. Undoubtedly, AI is influencing millions of lives at an accelerating pace (IBM Research, 2019; West et al., 2019).

At this point, it is oftentimes hard to draw the line between AI and buzzwords used like Machine Learning (ML), Deep Learning, Intelligent Systems, and Predictive Algorithm. According to Professor John McCarthy (as cited in JMC Stanford, n.d.), AI is the “science and engineering of making intelligent machines, especially intelligent computer programs”. Due to AI, machines can learn from past experience and can perform human-like decisions and tasks. No matter if AI is in the form of a self-driving car, movie suggestion system, or a chess-playing computer, these systems rely extremely on ML, deep learning, and natural language processing (SAS, n.d.). Thereby, a machine carries out a task built upon a set of prescribed rules that solve problems. This “intelligent” concept of an algorithm is what can be understood as AI. ML is, therefore, a subset of AI, empowering systems to learn from data. ML contains another subset; deep learning which is the next evolution from ML inspired by the human brain’s information processing patterns (Garbade, 2018).

AI systems nowadays change not only people's lives, AI also has huge impacts on businesses, and therefore, the world economy as a whole. According to a McKinsey study (Chui & Malhotra, 2018), 47% of business executives claim to be using at least one AI capability in their organization processes. Furthermore, 21% say their business embedded AI in multiple processes and 30% are now exploring a pilot of an AI system. Another McKinsey study (Cheatham, Javanmardian, & Samandari, 2019) shows that almost 80% of the interviewed companies experience values from their deployed AI systems. By 2030, however, 70% of global companies will adopt at least one kind of AI technology. This adoption will result in an additional global economic value of 13 trillion U.S. dollars which is equivalent to a 16% higher GDP in 2030 compared to 2018. This gain is likely to be related with the increased labor automation and innovation resulting in increased productivity and efficiency (Bughin, Seong, Manyika, Chui, & Joshi, 2018). As an S-shaped AI adoption curve (meaning first a flat curve, followed by a steep increase and then a static curve progression) is most likely to hold true, research assumes that the AI contribution to the world economy will build up at an accelerating pace with time (Accenture, 2017; Bughin et al., 2018). Especially the United States of America (U.S.) and China will contribute to this success as they are the pioneers in AI research, development, and deployment. In particular, China is catching up with the U.S. as Chinese internet companies like Alibaba, Baidu, and Tencent are investing billions in AI technologies (Charlton, 2019). While the U.S. puts emphasis on research and development in the field of ML natural language processing, neural networks, and recommendation engines, China focuses on sensors, intelligent robotics, and predictive maintenance. Other global players also found their niche, like South Korea focusing on intelligent mobile terminals and display reality, as well as Germany, being strong in the field of gesture recognition, augmented reality, and vehicle control. As each country is intensifying its AI research and development in certain sectors, global overreaching coverage of AI systems into all possible areas is reasonable (Accenture, 2017).

As AI boasts promising values with respect to individuals, businesses, and the world economy, AI has its downsides that have recently come to light. Whereas AI development will surely benefit numerous individuals and corporations, some will be disadvantaged by these same systems.

First, AI is increasingly applied by some developed countries, hence it is not so prevalent in developing ones. Such imbalance can potentially trigger the risk of further enlarging the existing economic gaps. The same goes for business, Accenture (2017) predicts that the AI landscape will be mostly dominated by a few large payers, causing difficulties for small businesses and resulting in a narrow, unequal, and unhealthy landscape. Similarly, immense changes in the workforce have been forecasted, creating new opportunities for a few, and leaving multiple workers behind, which is especially true when it comes to low-paid jobs. AI is predicted to replace multiple repeatable tasks that have been filled by humans for whom new occupations have to be created (Forbes Insights & Intel AI, 2019).

Second, the AI ecosystem causes concern for data security and privacy. The typical need for AI systems to have tremendous amounts of data makes more and more companies collect, store, process, and potentially even sell customers data. Data has become the "new gold" and incrementally more companies realize and leverage its potential. In the era of AI, data protection regulations are needed. However, with the rise of AI, the concept of privacy has "changed over time" and has become "increasingly complex", abstract and hard to grasp, says Professor Bernhard Debatin, the director of the Institute for Applied and Professional Ethics at the Ohio State University

(as cited in Forbes Insights & Intel AI, 2019, para. 6). The increasing availability of human behavior data allows training more complex AI models which will have an increased potential to identify individuals in a crowd, accurately forecast peoples' actions, and predict future behaviors. This opens a new door for targeting of interventions, opinion manipulation, and eventually limitations of freedom of action.

Finally, a threat that has received further attention within the last few years is the risk of AI systems being biased. "The real safety question (...) is that if we give these systems biased data, they will be biased," said Giannandrea before a conference at Google on the relation between humans and AI (Knight, 2017b, para. 2). Biased AI systems are already pervasive in many aspects and areas of our lives. As more than ever, important decisions are taken over by AI systems, the more severe are the consequences. The AI system is invading sectors such as healthcare, employment, criminal, and finance amongst others. Therefore, life-changing decisions are potentially based on such biased systems. For instance, such decisions can include the invitation for a job interview, granted parole on criminal trials and convictions, detection of skin cancer in clinical research, or the granting of mortgage or student loans. The impact of the biased outcomes will become even more severe as the technology is spreading to more critical areas and more complex algorithms. Thereby, AI systems are being used by individuals who do not necessarily have a deeper technical understanding of AI models and thus, are not able to scrutinize the systems. Especially those living in poorer communities and/or belonging to minority groups are mistreated by biased systems, as they are discriminated against by prejudices encoded in the data. Consequently, the AI system that was previously perceived as objective actually causes severe discrimination and unfairness on the basis of human errors. However, even if the severe consequences of biased AI are known, almost no companies or governments feel responsible to focus on addressing such problems (Knight, 2017a, 2017b).

In this dissertation the focus lies on the latter problem of biased AI. Our aim is to profoundly understand the source of the problem as well as providing insights to decrease biases in AI systems and help teams of individuals to be more aware of the challenges they face when developing AI systems for social applications.

## 3.2. BIAS & STEREOTYPING

A bias can generally be understood as an outcome that is systematically off the mark from its reality and that, thus, contributes to unfairness or imprecise perception (Kozyrkov, 2019). Over 180 human biases have been identified to influence how humans behave and make decisions. A bias can be revealed through an unfair representation of a person based on their social groups such as gender, race, age, education, income level, and sexual orientation (IBM Research, 2019). In practice, biases are common in mental-processing tasks such as heuristic estimation, target recognition, and social judgment.

Once a human has formed its opinion that can be based on beliefs, general acceptance, or likeability, the individual will process new incoming information in a way that is in line with the person's opinion. To change a deeply entrenched view depends on whether individuals approve or refute it (Tobena, Marks & Dar, 1999). In order to overthrow a strong belief or attitude, massive contrary evidence is required and even the greatest amount can sometimes be insufficient. Already in 1876, Darwin (as cited in Tobena et al., 1999) realized that he was more prone to forgetting observations and thoughts contradicting his hypothesis than those that favor it. A more recent study (Klayman &

Ha, 1987) confirms his observation. Humans tend to cherry-pick evidence supporting what favors their hypothesis and simultaneously neglect disconfirming evidence. A selective tendency is the consequence of weighting options in our cognitive routines. Such a weighting normally aims to cause efficiency; however, it can also provoke false observations and subjective reasoning (Tobena et al., 1999). Many errors in judgment emerge from overreliance on heuristics, also referred to as cognitive shortcuts. These shortcuts can entail evaluation based on similarities of previous events or processes. Oftentimes, when information is missing or time is insufficient, cognition relies on these heuristics (Baron, 2000; Epstein, 1994). As a result, relevant information can be overlooked and distorted, hence causing adulterated outcomes and results. Certainly, these neurocognitive systems' characteristics can result in inefficient performance and human bias (Tobena et al., 1999).

Social stereotyping, as well as social categorization, has a crucial stake in supporting prejudice and social conflict. Social stereotypes reinforced by biased cognitive processing are generally formed easily (Tobena et al., 1999). As an example, a single negative interaction with a Black American can result in stereotypical thinking (Henderson-King & Nisbett, 1996). These quickly formed stereotypes affect people's perception of others, and therefore, also their behavior and actions (Tobena et al., 1999). Information about people is memorized in a way that is in line with the stereotype (Gergen & Gergen, 1985) and additionally, negative information is better recalled than positive (Howard & Rothbart, 1980). Individuals believe in "illusory correlations" (Chapman & Chapman, 1969) which contribute to the "illusion of validity" of their own beliefs and view of the world and others (Kahneman, Slovic, Slovic, & Tversky, 1982). In other words, individuals perceive covariations which actually do not exist but still give them an incorrect sense of certainty. The underlying reason is that time constraints often do not allow people to wait to be fully and thoroughly informed in order to form an opinion about their counterparts. To save time, individuals often subconsciously create a snap judgment based on their established stereotypes. This quick categorization is based on generalizations that speed up the decision and evaluation process that would have required to take into consideration multiple attributes. For instance, judgments about honesty, loyalty, or animosity are formed quickly without comprehensive consideration (Tobena et al., 1999; Ridley, 1996).

Similarly, ingroup biases also emerge easily and do not need to be explicitly trained (Tajfel, 2010). People quickly form social groups such as "friends" and "enemies", whereby a single label can be sufficient to make such a biased categorization. Ingroup bias has, therefore, a self-serving purpose as individuals have the deep need to find their spot inside a group (Wilson & Sober, 1994). However, the costs paid for this mental processing shortcut are high; prejudice against certain individuals and groups can cause potential discrimination (Tobena et al., 1999).

## 3.3. DISCRIMINATION

### 3.3.1. Disparate Treatment

Disparate treatment designates direct or intended discrimination. Thereby, individuals or groups are being discriminated against based on explicit choices made on the basis of their belonging to a certain group. Disparate Treatment can be understood as the raw form of discrimination wherefore it is essential to understand its nature (Feldman, Friedler, Moeller, Scheidegger, & Venkatasubramanian, 2015).

According to the Cambridge Dictionary (n.d.-b), discrimination is “treating a person or particular group of people differently, especially in a worse way from the way in which you treat other people, because of their skin color, sex, sexuality, etc.” The Equality Act that protects UK citizens from discrimination defines nine so called “protected” characteristics: age, disability, gender reassignment, marriage and civil partnership, pregnancy and maternity, race, religion or belief, sex, and sexual orientation (University of Sheffield, n.d.).

Having laws for protecting specific groups was not always the case. Women were placed worse in almost all countries of this globe, black people faced slavery especially in the U.S. and colonial countries, while Jews and Roma have suffered repeatedly from ethnic profiling and social persecution. These devastating occurrences have been followed by new laws in the respective countries to ensure more equality (e.g. Human Rights Campaign, 2019). Such structural adjustments in the law still do not ensure that racial bias reflects the new legal standards (Dovidio & Gaertner, 2004). Research shows that the principle of equality has thrived from the 1960s onwards, however, some level of inequality remains (Bobo, 2001). For instance, in 2018, the median African American household earned only 59 cents, and the median Hispanic household 73 cents on the dollar of a white American household (Wilson and Williams, 2019). Furthermore, in 2015, only 55.5% of non-Hispanic American blacks in comparison to 75.4 % of non-Hispanic white Americans are covered with private health insurance. Impacts of the gap in health insurance are that U.S. black Americans have a 3.7 shorter life expectancy than white Americans (U.S. Department of Health and Human Services Office of Minority Health, 2019). Additionally, social studies (Massey, 2001) indicate that blacks are oftentimes a segregated isolated group from the rest of society mostly caused by the corollary of discrimination.

As disparate treatment might be perceived as being more visible, techniques might have been applied to cover direct discrimination. One technique is reverse tokenism, whereby individuals of the majority class are intentionally assigned to the negative class. The purpose is to oppose the claim of discrimination by the protected class by referring to the rejected members of the majority class as a vindication. For instance, a rejection of a minority applicant and can be justified with the rejection of a white male applicant. Following the prominent ruling of the Ricci v DeStefano decision (Feldman et al., 2015; Supreme Court of the United States, 2009), reverse tokenism can be classified as disparate treatment as the majority group was intentionally rejected.

### 3.3.2. Disparate Impact

The U.S. Supreme Court ruled on the Griggs v. Duke Power case and thereby put forth the legal doctrine of disparate impact. Until today, the legal term is used to describe unintended or indirect discrimination (Feldman et al., 2015; Supreme Court of the United States, 1971). In the ruling, the Duke Power Co. was compelled to abandon intelligence test scores, diplomas, and qualifications that highly correlated with race in their hiring decisions. In the end, this case made up the foundation of the disparate impact and disparate treatment framework which is still valid and lays the ground for prosecution until today.

Direct discrimination is based on human prejudices and makes its contribution to inequality in today’s world. A great stake of direct discrimination is what sociologists define as “institutional” discrimination. Institutional discrimination is not actually based on intentional decisions but rather on the prejudice of prior decision-makers or the widespread biases anchored in society. An algorithm

can exhibit these biases even though they have not been explicitly programmed to entail any tensions. Due to disparate impact being commonly a result on unintended discrimination, justice is hard to enforce as the problem is oftentimes difficult to identify and address. Furthermore, discriminate impact can evolve in various steps in data mining such as target variable definition, data collection, data labeling, feature selection, and deciding based on the model's outcomes. Each step can potentially be the source of the disparate impact causing difficulties in identifying the root causes of the problem. Even if each step is carried out in an extremely careful manner, unintentional biases can remain as the algorithm can retrace the protected class based on its relation to proxy variables. The combination of the external biases and prejudices from individuals and society as well as the complications of identifying the root causes hampers the prevention of the disparate impact. In fact, research still fails to sufficiently provide issue resolutions to the underlying problem (Barocas & Selbst, 2016).

One attempt to provide a tool to quantify disparate impact has been developed by Feldman et al. (2015). They introduced a generalization of the 80 percent rule to detect and measure disparate impact in their research. The so-called "80% rule" is the probability of the majority group having a positive outcome divided by the probability that a minority individual has a positive result. This measure should be kept below or equal to 0.8 to avoid conviction of disparate impacts. Originally, this rule has its roots in the U.S. Equal Employment Opportunity Commission's (EEOC) Uniform Guidelines suggesting a "four-fifth" test to identify adverse impacts in discrimination cases. The EEOC guideline has been especially used in hiring decisions, promotions, and other related employment judgments regarding employees' race, sex, and ethnic groups. However, this rule should be used with caution. As whenever the ratio of minority to majority groups is small, the guideline is less demanding than with higher ratios (Sobol & Ellard, 1988).

Depending on the literature, different terms are used to identify disparate impact. For instance, redlining is oftentimes used as a synonym and was originally used as the term to describe implicit discrimination relying on the correlation between certain people's neighborhood and their race (Calders & Verwer, 2010; Dwork et al., 2012; Kamishima, Akaho, & Sakuma, 2011). Besides, Dwork et al. call disparate impact in their paper "Fairness Through Awareness" discrimination based on redundant encodings (Dwork, Hardt, Pitassi, Reingold, & Zemel, 2012). The term redundant encodings describe the instance when the variable of belonging to a protected group is not explicitly used but is encoded in other data. The data is then used to make decisions leading to discriminatory outcomes.

# 4. TYPES OF BIASES/ PROTECTED VARIABLES

## 4.1. RACIAL BIAS

Even though the juridical form of racial discrimination was forbidden since the civil rights movement in the 1960s, racial discrimination might be attenuated but still exist in minority groups' daily life (Feagin, 1991; Sellers & Shelton, 2003). Racial discrimination can be found in the workplace, education, health care, and courts in all parts of the earth (Kamishima et al., 2012). Besides, it is present on the internet and media influencing an unlimited quantity of people. Lately, the issue has received tremendous attention and aroused worldwide protests under the name of the Black Lives Matter movement after a police officer kept his knee on the neck of the subsequently deceased African American George Floyd (Cheung, 2020).

Racism is an ideology that has its roots in prejudices which involves discriminatory behavior towards individuals due to their supposed "inferiority". This social belief relies on the idea that the human species can be distinguished into different human races that vary in physical and psychological aspects. However, scientific research exhibits that "human populations are not unambiguous, clearly demarcated, biologically distinct groups" (American Anthropological Association, 1998) and that the notion of race is purely embedded in individual imagination and social constructions. As all humans are part of the same species, it is nonsensical to differentiate between different races. Even though racism is based on a delusional construct that human nature can be categorized, the racist ideologies have been devastating humanity for centuries. It has been the cause and justification for slavery, colonialism, forced sterilizations, and mass murder. Unfortunately, racism endures in our society although mostly in diluted form.

Xenophobia and racism are often cited in the same breath, whereby Xenophobia is the "strong feeling of dislike or fear of people from other countries" (Oxford Learner's Dictionaries., n.d., para. 2). In other words, it is the irrational aversion to foreigners without any rational justification. This fear is solely a prejudice and is often not related to concrete negative experiences with foreign people. Individuals with Xenophobia have the false notion that individuals from different countries, cultures, groups, speaking different languages, pose a threat. Generally speaking, the more different the individual is perceived, the stronger the negative feeling of the person having Xenophobia. Xenophobia while closely related to racism (Sellers & Shelton, 2003), is different from racism. For instance, a U.S. study (Black, Schweitzer, & Mandell, 1978) shows that Afro Americans are less likely to get accepted for a mortgage lending than any other group including white, Puerto Rican, Mexican American other Hispanic, American Indian.

Racial discrimination can happen anywhere, however, it is probably most severe in legal courts. In fact, multiple studies have shown that dark-skinned convicts receive more severe penalties than white perpetrators (Steffensmeier, Ulmer, & Kramer, 1998; Petersilia, 1983; Spohn, 1990). Some researchers argue that racial discrimination in criminal courts has declined (Steffensmeier et al., 1998; Kleck, 1981) while other scholars (Steffensmeier et al., 1998) believe that nothing has changed over the course of time. However, these findings vary when analyzed in combination with other attributes such as age or gender. Especially young black men are being categorized as the "dangerous" group, argues Gibbs (1988) and are oftentimes portrayed by the media as being dangerous and dysfunctional. In particular research from before the turn of the millennium argues

that young black males are oftentimes associated with drugs, social welfare recipients and dropouts (Gibbs, 1988; Steffensmeier et al., 1998). Even if those views might seem outdated to a modern, educated human being, old views can still be reflected in data.

Besides, research has found several relationships between sexist, racist, and homonegative perceptions (Ficarrotto, 1990) meaning that if an individual is discriminating against others on the grounds of their minority group membership, they are most likely prejudiced against other minority groups too. This explains why especially black women suffer with biased AI systems, as they belong simultaneously to two minority groups.

## 4.2. GENDER BIAS

According to the World Bank report Women, Business, and Law 2019, only six countries globally provide equal legal work rights for women and men (World Bank Group, 2019). The measurement used in this study includes the ability for women to make economic decisions, employability, legal hurdles towards entrepreneurship, and pensions. According to the report, gender discrimination subsists in 187 countries. This relatively high number already incorporates progress, as a decade ago no country had equal legal rights implemented. The little advance made can be especially found in the parental leave legislation as well as the female political representation (Norris, 1987; Rule, 1981; Rosenbluth, Salmond, & Thies, 2006). South Asia made the biggest progress while countries in the Middle East and North Africa had the least improvement. The United Nations women report emphasizes empowerment in the workforce and argues that “when more women work, economies grow” (United Nations Women, 2018). In 2017, women had higher unemployment rates than men, 6.2 percent, and 5.5 percent respectively. In fact, if female employment rates in the OECD countries would match those of Sweden, the country with the highest female employment rate, the GDP would increase by more than 6 trillion USD (Eurostat, 2020; PwC, 2018). Adding to the controversy, other scholars (Cuberes & Teignier, 2016) state that failing to close the gender gap creates costs to the economy of roughly 15 percent of the GDP, whereby 40% is caused due to entrepreneurial gaps.

For centuries, the workforce and especially leadership positions have been exclusively filled with men (Eagly & Johannesen-Schmidt, 2001). Even though some time has passed since the publication of the study, women still suffer from gender discrimination in the workforce (Locke, 2019). Thereby, women do not enjoy the same treatment as men when it comes to hiring, firing, promotions, and wages. Studies show that male applicants are generally more positively evaluated than female applicants who have the same qualification. Even when the only difference is the applicants’ gender, males are more easily invited to interviews and offered a job.

One source of this problem lies in the roots of our social beliefs about male and female and leadership. Thirty years have passed since women became 50% of the college graduates in the U.S., and still, men fill most leadership positions in politics and business today (Sandberg, 2015). Common beliefs remain that men are supposed to be “agentic” meaning decisive, strong, and assertive, while women should be communal in other words warm, caring, and sympathetic. In perspective, the stereotypical leader is characterized by being self-reliant, dominant, assertive, and competitive. Past research found a positive relationship between dominance, masculinity, and leadership (Mann, 1959; Zaccaro, Kemp, & Bader, 2004). Because men have held leadership roles over the majority of recorded human history, their style has been the accepted perception of an optimal leader (Eagly & Johannesen-Schmidt, 2001). Consequently, the social understanding of a leader is still more in line

with the characteristics of men rather than women. As a result, women have to behave counter-stereotypically in order to be seen as a leader (Eagly & Johannesen-Schmidt, 2001; Locke, 2019). However, even if women and men have the same personality traits as a leader, men are generally more liked. Reasons to believe are provided by Professor Frank Flynn (as cited in Katsarou, 2019.; Sandberg, 2015) who presented a case study to his class at Columbia Business School whereby students were given a case about Heidi Roizen's professional success story as well as a modified version with the male name "Howard". Students believed that Heidi was substantially less likely and was perceived as selfish. In contrast, the male" was perceived as authoritative and dominant which are positive leadership traits (Sandberg, 2015).

Besides, several studies have shown that women have the advantage of being more cooperative and collaborative which is actually a suitable leadership trait that has just not been accepted as being one by society (Book & Book, 2000; Locke, 2019; Rudman & Kilanski, 2000). Still, women are often evaluated more negatively than comparable male employees (Davison & Burke, 2000). Some still perceive the social role of women to care for children and the household while men's role is to financially support. (Eagly, 2013; Nadler & Kufahl, 2014). This social cognition of the different gender roles has been partly responsible for gender bias in the workforce (Eagly & Karau, 2002).

Gender discrimination does not stop beyond the borders of financial institutions. More than 1.3 billion women have not had any contact with the financial system or services (Demirguc-Kunt, Klapper, & Singer, 2013; see also The Economist, 2013). Research by the World Bank (Demirguc-Kunt, Klapper, & Singer, 2013) showed that women are highly less likely to have a bank account, save and borrow in countries with greater legal discrimination. Access to financial systems is essential to achieve equality as it opens the doors to monetary independence, asset ownership and it promotes economic growth. A study of credit outcomes of a large Albanian bank (Beck, Behr, & Madestam, 2018) shows that women who were assigned male loan officers are more likely to pay higher interest rates with shorter maturity and receive lower loan amounts. Findings indicate that women might obtain lower credit scores, which are highly crucial for modern lending technologies (Narain, 2009). Another study (Black, et al.,1978) revealed that females are less likely to be granted a mortgage loan at all.

In most scenarios, men are given preferential treatment over women. However, their position is worse in the criminal sector. Research agrees upon the finding that in prison sentencing, female defendants receive milder penalties than males (Bickle & Peterson, 1991; Daly and Bordt, 1995; Steffensmeier et al., 1998; Steffensmeier, Kramer, & Steifel, 1993). However, when put in combination with the race factor, Klein and Kress (1976) argue that race discrimination is greater among women than men. Therefore, white females will get much lighter sentencing in contrast to black females who receive sentences similar to white men. Steffensmeier et al. (1998) compared the effect of the minority group membership on sentence outcome and found that the gender variable had the largest influence on the sentencing, beating age, and race.

In today's world, gender is not two-sided (either males or females), but several transgender types exist. Buolamwini and Gebru (2018) argue that all companies they analyzed use as gender classification features binary sex labels which were either female or male. Making people choose for instance in an application process between a female and male category is therefore not in line with social and psychological standards that acknowledge a more continuous and fluid approach to

gender. In addition, a study by the University of Colorado Boulder found that common facial analysis tools like Amazon's Recognition, Microsoft's Azure, IBM's Watson, and Clarifai incessantly misidentify people with non-binary gender. While cisgender men and women enjoy accuracy rates of 98%, trans men were wrongly categorized at around 30% during the trial. The tools performed much worse for non-binary or genderqueer people (Khalid, 2019).

### 4.3. AGE BIAS

The Age Discrimination in Employment Act (ADEA) protects people above 40 years old against unfair treatment in the workforce including hiring, firing, and payment, job assignments, layoffs, promotions, training, benefits, and other working conditions. In Europe, the European Union put a directive in effect enforcing equal treatment in the employment of all age groups. However, several studies show that older employees are oftentimes perceived as being slower, less adaptable, less creative, less open towards training, and prone to health issues (Aiyar & Ebeke, 2017; Finkelstein & Burke, 1998; Rhodes, 1983). Even if against the law, age discrimination exists illustrated by the thousands of legal actions taken on the basis of the Age Discrimination in Employment Act each year in the U.S. (Finkelstein, Burke & Raju, 1995). The research shows that especially younger people have stereotypical views towards older people in contrast to old people having fewer negative attitudes towards younger employees. A reason research proposed that older people experienced all age levels already and are, therefore, more sympathetic. As a result, young people are more likely to hinder old people from being hired or promoted.

The meta-analytic test of Finkelstein et al. (1995) shows that younger workers between the ages of 17 to 29 years perceive younger workers generally as more favorably than people above 30 years old. This research is in the line provided by several other scholars (Craft, Doctors, Shkop, & Benecki, 1979; Levin, 1988). In contrast, older employees did rate young and older employees equally. The conclusion can be drawn, that older people less heavily discriminate on the basis of age as in the case of younger employees.

As age discrimination is present in the employment decisions, so it is in the financial sector. A study on mortgage lending (Black et al., 1978) shows that that the acceptance rate increases as the age increases and then falls again. In the criminal defense sector, age discrimination becomes more complex. At first sight, when comparing "old" (above 50 years old) vs "young" (between 18 to 50 years old), the older group is treated more leniently than the younger ones (Champion, 1987; Cutshall and Adams, 1983; Steffensmeier et al., 1998). However, when categorized in a more detailed manner, meaning aged youth 18-20, young adults aged 21-29, adults 30-50, and elderly people above 50, an inverted U-shaped relationship can be found (Steffensmeier et al., 1998; Steffensmeier, Kramer, & Ulmer, 1995). In fact, the inverted U-shaped curve has a low between the ages of 18 to 20 as judges perceive them as still formable, less dangerous for the community, and generally more sympathetic. The peak exists in the young adults' group, who receive the harshest sentencing. From age 30 on the sentencing harshness linearly declines along with age. Instances like these show that age discrimination exists in all kinds of areas and when encoded in data can be perpetuated by biased AI models.

## 4.4. Sexual Orientation Bias

The European Union created legislation that protects people from discrimination on the basis of sexual orientation just in the field of employment. Solely eight member states extend the laws to cover the rights beyond employment which are defined in the Racial Equality Directive (De Shutter et al., 2009). Only 20 of the 50 U.S. states cover sexual orientation discrimination in their laws and protect these people from unfairness in employment, housing, and public accommodations (Freedom for All Americans, n.d.). Anyhow, the LGBT survey of 2013 by the Fundamental Rights Agency reveals that 47% of the 93,097 respondents experienced discrimination due to their sexual orientation within the last year (Bell, 2017; European Union Agency for Fundamental Rights, n.d).

Homonegativity defines the negative affect, behavior, and cognitions towards people who are recognized as being - rightly or wrongly - lesbian or gay. Morrison and co-scholars (Morrison & Morrison, 2003, 2011; Morrison, Morrison, & Franklin, 2009) distinguish two types of homonegativity: old-fashioned and modern homonegativity. Old-fashioned homonegativity is denoted by morals or religion which objects homosexuality whereby modern homonegativity reflects contemporary concerns. Concerns contain the view that homosexual people change their current status, and therefore, affect current settings (Morrison & Morrison, 2011). A report by the Council of Europe states that a negative attitude towards homosexual people is mostly shaped by biased, outdated, and wrong information which creates an incorrect stereotypical portrayal of these people. Furthermore, they invoke that LGBT persons are still being discriminated against at work and school.

Roughly 25% to 66% of homosexual female and male employees and 75% of transsexual employees experienced some kind of discrimination in the workforce (Nadler & Kufahl, 2014). Even if policies exist, individuals fear to reveal their sexual orientation due to their fear of ostracism, hostility, fewer promotions or even being fired (Ragins & Cornwell, 2001). In fact, several studies show that gay and bisexual oriented men earn comparability less than equally qualified heterosexual men. In contrast, lesbian and bisexual women do have equal or higher wages than comparably qualified heterosexual female counterparts (Badgett & Frank, 2007; Black, Makar, Sanders, & Taylor, 2003; Blandford, 2000;). A study by Black et al. (2003) reported findings of wages between 15% to 16% lower for homosexual men than for heterosexual men and between 20% to 34% higher for homosexual women than for heterosexual women. This advantage of lesbians is in line with the findings of section 4.2. that stereotypical female gender roles are not appreciated in the workforce and that not standardized female manifestation will, therefore, result in higher earning (Weichselbaumer, 2003). Lesbians are oftentimes perceived as being more masculine, dominant, independent, assertive, and detached corresponding with labor market success criteria ordinarily aligned to male recipients (Blandford, 2000; Clain & Leppel, 2001; Reiss, Safer & Yotive, 1976; Weichselbaumer, 2003). Furthermore, the oftentimes lack of children might lead to higher investment in their training and prospects for higher positions (Weichselbaumer, 2003). Still, 16 to 46 percent of gay and lesbian respondents of a survey reported some kind of discriminating experience in the workforce (Badgett, 1995; Drydakis, 2009). The revealed higher wages of lesbians can be put in proportion to their productivity and characteristical differences, however, discrimination remains. In Europe, sexual minority individuals have been repeatedly argued to have been fired, not promoted, not hired due to their sexual orientation (European Union Agency for Fundamental Right, 2009).

In the health insurance sector, sexual minorities are more likely to not have health insurance and health care access when compared with heterosexuals (Badgett & Schneebaum, 2015; Buchmueller & Carpenter, 2010; Charlton et al., 2018). Six percent of the lesbian, gay, bisexual, and queer respondents of a study and 12% of transgender people reported having been denied healthcare due to their actual or sexual orientation or gender transition (Mirza & Rooney, 2018).

Even though laws attempt to protect people with a sexual orientation other than heterosexuality, they often fail to do so. Biased decisions made based on sexual orientation can potentially be fed into the system and can maintain this bias in the future.

### 4.5. Disability Bias

According to the U.S. Census Bureau, approximately every fifth person living in the United States has some sort of disability (Brault, 2008; Scherer, 2017). In this paper, the term disability is understood as by the United Nations (2007, para. 5): "an evolving concept and … results from the interaction between persons with impairments and attitudinal and environmental barriers that hinders their full and effective participation in society on an equal basis with others."

The American with Disabilities Act (ADA) of 1990 outlaws discrimination on the basis of disability in the United States. For instance, it prohibits discrimination in employment, government programs, and services, public transportation, and accommodations. In Europe, disability is acknowledged as a ground for discrimination. Furthermore, the Framework Directive secures workers and candidates from discrimination in occupation and employment in Europe. Moreover, 177 countries have signed the Convention on the Rights of People with Disabilities which furthermore extensively protects people with disabilities. However, scholars examining the topic assess the severe impact on people with disabilities by the rise of AI tools (Scherer, 2017; Trewin, 2018).

In contrast to other discriminating variables, disability is hard to categorize into a variable. Especially, the different, nuanced and unique types of disabilities differentiate itself from other protected groups. Disabilities can have several dimensions and individuals can experience several disabilities at the same time. Furthermore, disabilities differ in terms of intensity and impact on individuals. In addition, intensity, impact, and conditions can change as time goes by, which can be understood as another hurdle to classify adequately. Even within groups that seem to be homogenous, different types exist as the disabled community is dominated by outliers (Trewin, 2018).

As machine learning works by finding patterns in datasets, varieties in disabilities pose a challenge to categorize efficiently. Machine learning bases its learning on homogenous groups and predicts based on similar characteristics. Outliers are thereby often eliminated from the datasets as it is regarded as "noise". As the disability community includes many individual types of disabilities, systems can wrongly identify them as outliers. In contrast, when models include outlier data in their training, the model becomes more complex and more likely overfits the data. Therefore, simpler models are generally preferred. However, the system can be unable to identify a pattern and create unsuitable outputs for people with disabilities. The outputs can be of poorer quality or unfair compared to other individuals. Attempts to solve the problem of bias towards the other protected groups do oftentimes include gathering a balanced training data set. The nature of disabilities impedes to address the problem likewise and creates a need to find another solution for disabled people (Trewin, 2018).

The solution approach “Fairness Through Unawareness” is proposed whereby no information about the disability is included in the decision-making process. For instance, many applicants do not reveal their disabilities in job applications as the ADA prohibits employers from demanding information about the applicant’s disabilities. However, the correlation between other attributes and disability can indicate their limitations indirectly. For example, the need for assistive technologies during an online test such as screen readers or magnifiers can point out disability.

Nonetheless, it is crucial to include disability information to detect any systematic discrimination. When the disability data is not given it is difficult to estimate if people with disabilities have more negative outcomes than people without disabilities. AI scholars like Dwork et al. (2012) have advocated the “fairness through awareness” concept, whereby the protected group is intentionally known that fairness can be formally specified, tested, and forced algorithmically. Numerical methods can then be used whereby the output of models is adjusted to mitigate bias. Besides, several models (Gajane & Pechenizkiy, 2017; Verma & Rubin, 2018) and toolkits (Bellamy et al., 2018) can be applied to assess accuracy, fairness, and bias and can evaluate the effectiveness of bias mitigation interventions.

As a case in point, a recent field study detected that including disability information in a job application can conclude in 26% less positive responses even if the person’s disability did not negatively impact work productivity (Ameri et al., 2018). Besides, another study disclosed a pay gap between people with and without disability which cannot be justified with performance differences (Kruse, Schur, Rogers, & Ameri, 2018). With this in mind, it is understandable that people with disabilities oftentimes wish to not share information about their health condition and limitations.

Privacy concerns are another crucial factor when it comes to sharing disability information to employers even though it could help to establish fairness. Their wish to have the control and knowledge about how their disability information is used can oftentimes not be guaranteed. In fact, it has been disclosed that the ACT standardized testing organization in the United States transmitted disability data of applicants to colleges and even sold the information to third parties (Jaschik, 2018; see also Bloom v. ACT, Inc., 2018).

Another concern is algorithmic fairness in the respect that the models work equally well for people with and without disabilities. As AI tools are already performing worse for different genders and races as a result of underrepresentation in the datasets, the representation is even worse for people with disabilities. Especially, speech recognition and facial recognition tools can underperform as the individuals’ speech, appearance, and behavior deviate from the norm. In this respect, it can be helpful to use a more diverse training data set or even build customized and specialized models for known groups. For instance, deaf speech is oftentimes not understood by speech recognition software which must be adjusted for these special groups.

Generally speaking, gender, racial and age bias in AI has received much more attention than disability discrimination in research and publications. Individual academics highlights the danger of applying the typical training and assessment methods for the machine learning models as they oftentimes do not apply to people with disabilities. One can make a safe assumption that more extensive and solution-oriented research is required in this field; however, it is beyond the scope of this thesis.

# 5. AREAS WITH BIASED AI

## 5.1. EMPLOYMENT

One of the most pernicious effects of AI can be seen in the employment sector (Ajunwa, Friedler, Scheidegger, & Venkatasubramanian, 2016). Nowadays, AI algorithms decide who gets employed, who gets fired, and how high wages and bonuses are set. Thereby, numerous demographic groups such as women, ethnic and national minorities, disabled people, religious and sexual minorities face unfair negative market outcomes (Weichselbaumer, 2003). How AI is evolving these processes and how severe these systems can aggravate people's life at the same time will be the center point of this part.

### 5.1.1. Hiring

As companies apprehend the increasing return on investment (ROI) from hiring the right person for a job, AI is disrupting the whole area of recruitment and identifying talents. According to McKinsey (Keller & Meaney, 2017), an over-performing person, the so-called "star", can produce more than 800% value than an average worker. These star performers decrease costs as their productivity rates can replace various workers. Another Harvard Business School study (Housman & Minor, 2015) illustrates the other side of the performance curve: the toxic workers. Toxic workers can drastically damage the performance of the organization as they engage in bad behavior negatively affecting their co-workers (Pierce & Balasubramanian, 2015). Negative effects are monetary costs, employee turnover efforts, customer loss, employee's moral loss, and damage to the legitimacy towards external stakeholders (Housman & Minor, 2015). In summary, hiring decisions are essential for organizations' performance and should be executed in a careful and optimized manner.

In contrast, Chamorro-Premuzic (2019) claims human hiring decisions to be intuitive and ineffective as decisions are mostly based on meritocratic criteria. In his opinion, AI-based hiring tools are the answer to human inefficiencies as they can reduce errors, noise, and bias. He concludes that women's generally inferior chances in the workforce would enhance. As men have been employed primarily because of their confidence with less regard to their competence, he argues, women would benefit from tools including relevant factors such as emotional intelligence, self-awareness, and coachability. As AI has its advantage of detecting important data points, so are its benefits of not ignoring aspects. An unbiased AI system, Chamorro-Premuzic (2019) states, can decide regarding the applicant's ethics, gender, or sexuality and can therefore in an optimum manner decide completely objectively. Even the most ethical, well-intentioned, and tolerant recruiter cannot completely ignore these characteristics. The more a person strives to ignore these notions, the more present they will be in his head as the human mind tends to over-compensate. Furthermore, humans are generally overconfident in their ability and will be most likely incapable to notice their partiality or bias and even less likely admit to it (Chamorro-Premuzic, 2019; Kahnemann, 2011). Unlike humans, AI does not mind admitting mistakes and eventually learns from errors made.

However, the idea of an AI hiring system being unbiased is difficult to bring to reality as human biases are encoded in the data the model is trained with (Chamorro-Premuzic, 2019). For instance, like Amazon's hiring tool which was accused by Reuters in 2018 of having developed a model that was indirectly discriminating against women. As Amazon tripled its employees from June 2015 to

575,700 workers, they needed to automate part of their hiring process to cover the demand (Dastin, 2018). Amazon created a hiring tool that learned from past experiences from the last 10 years and efficiently chose qualified applicants by comparing these to the characteristics of previous hires. Within this time period, most hired candidates were men as they were dominating the tech industry. Thereby, the word “women’s” in the applications was downgrading the applicants’ rank, and therefore, unintentionally decreasing female applicants' chance of getting hired. Even the digital leader Amazon could not guarantee an unbiased machine; thus, eventually banned the algorithm completely (Dastin, 2018; West et al., 2019).

In the United States, around 55 percent of human resources managers state that AI will be applied in their work within the next five years (Dastin, 2018; Career Builder, 2017). For instance, large players like Hilton Worldwide Holdings Inc and Goldman Sachs Group Inc. have been already exploring to automate portions of their hiring processes (Dastin, 2018). Goldman Sachs created a resume analysis tool that declares candidates as “best fit” according to the company’s standards and demands. Besides, the world’s largest professional network LinkedIn provides algorithm rankings of prospective candidates to employers based on the fit to their job postings. However, vice president of LinkedIn Talent Solution John Jersin argues that their services do not replace human recruiters. He states that he would not trust any AI system at this state to make any hiring decision by itself as the technology is not ready in his opinion (Dastin, 2018). Furthermore, Kevin Parker, the chief executive of HireVue, said that his firm analyzes speech and facial expressions of their interviewees in video interviews to rely less on their resumes. As stated in this paper before, some minorities face lower accuracy rates in facial recognition software which has to be explored if they face these rates in these applied tools as well.

Discrimination in the employment sector starts, however, long before the selection of the candidates. A recent study by the Carnegie Mellon University discovered that men were shown more high-paying executive jobs than women (Datta, Tschantz & Datta, 2015). Again, this study disclosed the problem of biased AI without complementing it with further research on how to straighten out the error. Moreover, a 2016 class-action lawsuit accused Facebook Business tools of providing a platform to discriminate African Americans, Latinos, and Asian Americans by excluding them from advertisements for relevant opportunities as it enables users to choose demographics similar to existing workers, thus reinforcing existing demographic disparities in the company. Under American federal law, employees can choose employees based on a “cultural fit” making this procedure not explicitly illegal. However, in this case, Facebook changed its ad platform to resolve the given legal dispute (Ajunwa, 2019). Instances like these can result in a closed-loop system that reinforces the bias. As these systems encourage certain people to apply to certain jobs these people do more likely apply and a fraction of those will then most likely get the job. The results will then be looped back and used to develop criteria for targeted job advertisement, hence, decides on who gets selected for a job (Dastin, 2018).

Rejected candidates are the ones suffering from biased hiring AI tools. However, they generally do not know if an AI system has been used in the recruitment process nor that AI bias was the cause of their negative outcome (Dastin, 2018). The system mostly lacks transparency, accountability and provides no platform for rejected candidates to get valid information nor provides tools to complain (Ajunwa, 2019). For example, a college student with an almost perfect Scholastic Aptitude Test (SAT) score and with a diagnosed bipolar disorder was rejected from multiple minimum-wage supermarket

jobs as he was requested to fill out a personality test (Ajunwa, 2019, O'Neil, 2016). The applied test detected his mental illness and thus he was rejected. He had to actively investigate the reason for his failure and initiate legal steps to challenge the received rejection. It can be assumed that this instance was not the first one of its kind and that not every applicant does have the mental capacity to challenge a negative outcome as well as the required time and financial resources to initiate a lawsuit.

In 2017, an investigation was initiated by the Illinois attorney general into several automated hiring platforms. A resume creation tool on Jobr did not provide applicants the option to choose a year of their college graduation or their first position before 1980. Consequently, older applicants were discriminated against (Ajunwa, 2019). As this discriminating flaw was visible to candidates it actually enabled victims to complain. However, more complex algorithm bias is hidden in the black box algorithm (see chapter 6.3.2.) which runs behind closed doors. If there is no claimant who identifies discrimination, there will be no judge to rule on it. Consequently, it can be concluded that the known cases of biased AI instances are just the tip of the iceberg and most instances are unknown.

For victims of biased AI systems to pass a federal lawsuit either proof of disparate treatment (proof of intentional discrimination) or disparate impact (statistical proof that a majority group was favored) is required (Ajunwa, 2019). As employers control the data of their hiring tools in platforms, it is extremely hard for applicants to get any kind of proof. Getting access to their own data is oftentimes not enough as they require other peoples' data to detect a statistical pattern to claim disparate impact or more specific algorithmic information about why their profile was rejected to prove disparate treatment (Ajunwa, 2019).

### 5.1.2. Firing

Biased AI in hiring received much attention in Biased AI research. However, other fields in Human Resources are also sensitive to biased AI tools. Similarly, firing has been mostly just named as being prone to AI bias, however, it has not been examined closely yet. O'Neil (2016) illustrates an example of an AI firing tool whereby teachers in Washington are assessed according to an IMPACT score whereby the bottom 2 percent is fired. The system incorporates bias, unfair measurements, and is intransparent for the affected teachers. Ranking tools as the one used in this example generally rely on hard facts but can oftentimes not consider the context and circumstances of the person's performance. Besides, the system did not generate any feedback control as teachers were simply fired and the system can therefore not be controlled and improved.

### 5.1.3. Compensation

Efficient incentive compensation management calls for high accuracy rates, well-grounded business decisions, and enhanced outcomes at the right time. AI technologies are also evolving in this field and provide opportunities for executive compensation decisions to be more efficient and economical. Cognitive compensation programs combine AI, advanced analytics, and machine learning to early identity problems, correct flawed schemes, and advance individual accountability (IBM Corporation, 2018). These systems monitor patterns in employees which either need to be reinforced with rewards or rectified when malicious action occurred. As the budget for Human Resources is generally limited, it is essential to allocate the monetary resources efficiently to develop and motivate the right workers (IBM Corporation, n.d-b.). Furthermore, these tools can generally

improve the experience of the employee as the compensation process aims to be more personally tailored to the individual employee (Guenole & Feinzig, 2018).

## 5.2. FINANCE

According to a study by PricewaterhouseCoopers (PwC, 2017), more than half of financial service corporations have made a "substantial investment" in AI and machine learning solutions. AI solutions can support banks and credit lenders to make important credit decisions based on an advanced analytical assessment of the lenders. An algorithm can analyze the risk and identify patterns by processing huge amounts of data points to identify early signs of potential problems.

Multiple firms are applying AI solutions to evaluate the creditworthiness and related risk already. For instance, the lending firm Crest Financial is using DataRobot, a machine learning software, to predict customers' likelihood of default and experience significant improvements (Bachinskiy, 2019; Schroer, 2019). Scienaptic Systems has multiple major credit card companies' clients providing a platform to learn from large amounts of unstructured and structured data. The Los Angeles based company ZestFinance developed the Zest Automated Machine Learning platform, a solution that supports companies to evaluate borrowers that have little or no credit history at all. They claim to cut their clients, an auto lender, losses by 23% annually by accurately predicting potential risks. Business examples like these illustrate how prevalent the implementation and use of AI systems is in the financial sector (Maruti Techlabs, n.d.; Schroer, 2019).

Besides, AI systems are also applied in fraud detection. The risk of credit card fraud has increased substantially due to the increase in e-commerce and related online transactions (Bachinskiy, 2019). Money laundering can also be prevented with the help of AI and can reduce the investigation workload by 20% (Ayasdi, 2018). Global financial players like Goldman Sachs and American Express have been applying this kind of technology, algorithms to identify unique patterns to detect fraud which can be updated in real-time (Bachinskiy, 2019).

Over the past five years, AI-driven investments have been rising and have closed in a trillion United States dollars in trading in 2018. Intelligent trading software processes huge piles of structured and unstructured data. The advantage lies in the speed with which this analysis and predictions can be done as time is essential in the financial sector. Most recently, Bloomberg launched an AI prediction matrix Alpaca Forest which predicts prices for investors. The AI system learns from real-time market data, identifies patterns, and thereon predicts price movements (Bachinskiy, 2019; Alpaca, n.d.).

With the benefits of AI systems, comes along the risk of biases. Especially, in granting common types of loans, mortgages, and student loans, AI systems provide a platform for bias and discrimination (Agarwal, Haq, & Coutinho, 2019). Nowadays, credit card companies have not just gathered basic account information, but store and eventually analyze the whole credit history and purchase behavior of their clients c, n.d.). According to a consumer payment study from 2018 (Total System Services Inc., 2019), 77% of consumers would rather pay with debit or credit cards than any other payment option. In the department store, just 7% and in supermarkets 13% of the respondents prefer paying cash. Consequently, the respective financial institutions know their clients' consumption patterns in a detailed manner. In other words, they know what they eat, where they dine, what they wear, where they refuel their car, and can also, most likely, pinpoint to where the credit cardholder works and lives.

As financial institutions have a wide range of customer data, the systems become more complex and less comprehensible. As the system learns from past data, human biases can be encoded in data, hence, discrimination can exist without being known. In Europe, for instance, male entrepreneurs are 5% more likely than female entrepreneurs to be granted a loan at a bank for their prospective business. In addition, these fewer women have been given the loan payment on average 0.5% higher interest rates on these loans. This gap cannot be explained by women's lower performances. In fact, the average technology company founded by a woman yields revenues 12% higher than the ones run by men (FinTech Futures, 2019). As data encodes these tendencies, women can still be discriminated against by these supposedly objective AI algorithms. Even if the protected class is not explicitly used in the algorithm, other variables can still correlate with the protected one and implicitly cause bias. For instance, a more female shopping behavior could lead to lower credit granting or higher interest as it implicitly learns from the previous decisions.

In China, customers are already purchasing goods by scanning their faces. It is unquestionable that facial payment technology will spread across China and will also find its way to the Western world. Customers must solely stop in front of a point-of-sale machine and have their face scanned which matches an image previously linked to their bank account (Agence France-Presse, 2019). Therefore, customers are no longer required to bring phones or wallets so that payment processes are executed faster and more efficiently than any other prior payment technology. Alipay, the financial payment provider to one of China's e-commerce leaders Alibaba, has announced to invest three billion yuan in AI technology within the next three years. Similarly, Tencent, the company owning WeChat with roughly 600 million users, also disclosed its new facial payment tool called "Frog Pro" (Agence France-Presse, 2019). In this early stage, no research exists regarding a biased facial payment system. However, since conventional facial recognition technologies have much lower accuracy rates for certain minorities like different races and gender the same lower accuracy rates can potentially apply to this new payment technology. Lower accuracy rates can result in security concerns that need to be addressed.

### 5.3. Criminal Defense

The United States has the largest criminal justice system in the world. In fact, more than 6.6 million people were under some type of correctional control in 2016 alone, including 2.2 million people who were detained in local, state, or federal prisons and jails. Namely, the United States is one of the world leaders of its imprisonment rate with almost 2.3 million inmates (Sawyer & Wagner, 2020; see also Collier, 2014; Kaebble, & Mary, 2018), surpassing the rates of almost all other nations (World Prison Brief, 2018).

The numbers, however, also comprise huge racial disparities dominating the U.S. criminal justice system. Especially, African Americans are targeted by these systems. Indeed, African Americans are 5.9 times more likely to be arrested than their white counterparts. Furthermore, Hispanics are 3.1 times as likely as white Americans to be incarcerated. Once arrested, African Americans have a higher risk of being convicted, and once they are convicted, they tend to face longer prison sentences (Carson, 2018). Research shows that every third black boy born in 2001 can anticipate going to prison one day, as can one of every six Latino boys. In contrast, just every seventeenth white child can expect to be imprisoned during his life (Mauer, 2011). Racial disparities among women also disseminated widely, however, not as substantial as for men (Sentencing Project, 2018).

Former Law professor and author, David Cole, describes in his book “No Equal Justice” the discriminating side of the U.S. justice system. In his mind, the double standards of the U.S. system are implicit as the criminal justice system is marked by laws that are color-blind and class blind. However, these laws contribute to the system's hazards as society perceives the system as fair and just refers to the theoretically fair laws when needed. As the system is in line with the constitutional rights, in theory, the Supreme Court can validate the outcomes of the criminal justice system as being fair. However, even if the system is formally fair, the prisons are disproportionately filled with poor black citizens, indicating otherwise (Puzzanchera, Sladky, & Kang, 2016).

### 5.3.1. Predictive Policing

With the rise in the amount and availability of data, comes the shift of reactive policing from solving crimes after they have occurred towards preemptive policing meaning predicting and preventing crimes in advance (Brayne, 2017; Hardyns & Rummens, 2018; Van Brakel, 2016). This is when predictive policing comes in place, using historical and real-time data to predict the time, location, and possibility of crimes and to identify the potential offender and victims of the underlying crime (Van Brakel, 2016). These predictive algorithms are based on assumptions such as the risk of re-offending and that certain social and economic conditions are prone to certain crimes (Brayne, Rosenblat & Boyd, 2015). Predictive policing can be subdivided into two types; predictive mapping and predictive identification. Predictive mapping is also known as “hotspot” policing and aims to predict locations prone to crimes for certain times. Predictive identification focuses on identifying certain people who are likely to become a perpetrator or victim of a certain crime (Jansen, 2018).

In Europe, data-driven policing is still in its infancy. However, legal mandates have been expanded to allow the police to collect more, retain longer, and share huge amounts of data (European Data Protection Supervisor, 2018, Jansen, 2018). To develop such algorithms, Europe has to improve the quality of the police data, wherefore, police officers are encouraged to enter respective data in the databases. Furthermore, the European Union invests heavily in the interoperability between databases, the building of new databases as well as in automating searches. Predictive mapping software is being developed and can be roughly subdivided into two types of models: near repeat and time-space. First includes the internal police data while the second also includes external variables like weather or events (Ferguson, 2019; Hardyns & Rummens, 2017). In Germany and Switzerland, several police departments use the German software PreCobs which is based on a near repeat model (Hardyns & Rummens, 2017; Van Brakel, 2016). The police forces of Amsterdam, North Rhine Westphalia, and Berlin all utilize a time-space policing model that calculates the possibility of crime occurring based on a broad spectrum of variables. Their models are called CAS, Skala, KirimPo, in the above order.

Predictive identification technologies are also utilized in Europe and help to identify potential offenders or victims of crime. In particular, risk modeling is applied whereby potential offenders are identified. For instance, the London Metropolitan University created a Gang Matrix that identifies potential gang members and their related risks to society. Research (Amnesty International, 2018; Scott, 2018) revealed, however, that the system especially targets young black men indicating a discriminatory pattern. In Amsterdam, two distinct risk modeling programs are being used; the top 600 and the top 400. The former identifies re-offenders who have been arrested for high impact crime in the last 5 years whereby the top 400 measures the likelihood of children below the age of 12

to eventually become criminals. The German Federal Criminal Police Office developed a system called RADA-iTE ranking the top 500 potential terrorists (Jansen, 2018).

In the United States, more and more police departments are using predictive crime software to better allocate their resources. Especially police departments where budgets are getting tighter are using this kind of software. The California-based start-up PredPol developed a patented algorithm that identifies times and locations where certain crimes are most likely to happen. Police departments can then target these locations and patrol to prevent crimes from occurring (PredPol, n.d.). The tool is currently used from departments in Los Angeles, Atlanta, and Alabama. Similarly, the city of New York uses CompStat and the Philadelphia police HunchLab (O'Neil, 2016). A predictive identification system is among others used by the Chicago Police Department whereby a list of approximately four hundred individuals is created that are most likely to commit a felony (O'Neil, 2016).

In particular, U.S. American Scholars have shown the potential risk of biases of these systems (Jansen, 2018). Especially black U.S. citizens are targeted like the disparity in imprisonment numbers between white Americans and minority group members illustrates. This disparity starts with the proportional higher contact of the minority groups with the police. For instance, U.S. surveys show that police are more likely to stop cars with black or Hispanic drivers than white drivers. Once pulled over, Hispanic and black drivers are three times more likely to be searched and two times more likely to be arrested than whites (Langton & Durose, 2013; Sentencing Project, 2018). Federal judge Shira Scheindlin denounced New York city's officials to have "turned a blind eye to the evidence that officers are conducting stops in a racially discriminatory manner," and that officers routinely "stopped blacks and Hispanics who would not have been stopped if they were white." In addition, she declared the city's stop-and-frisk technique as being unconstitutional in 2013.

O'Neil (2016) identifies the danger of predictive policing systems as they encode biases followed by severe unfairness. She describes the scenario that, because black people are more likely to undergo body searches, it is statistically more likely that the police will be successful during their proactive searches. Also, other scholars describe the over-policing of neighborhoods dominated by African American habitats (Jansen, 2018). Anyhow, the system gets reinforced as it learns from its successes. In addition, the systems target people in weaker economic positions, unable to afford good lawyers, and therefore, oftentimes charged with rigorous penalties. The system learns from all these incidents and creates a target that is young, male, and black. A vicious circle, or as O'Neil (2016) calls, a feedback loop is created, which spins and reinforces itself.

Discriminating predictive policing is not just prevalent in the United States. Some scholars identify that also European systems are marked by biases. In the U.K. systems, a negative feedback loop was found whereby black, Asian and other minority ethnic (BAME) are more likely to be arrested resulting in the system's assumption that the areas the BAME individuals live in are more prone to crimes (Couchman, 2019). The earlier explained Gang Matrix illustrates the systems' discriminating pattern as the information about the individuals categorized as gang members were shared with education, housing, and health care providers as well as with the driver and vehicle licensing authority (DVLA). Consequently, the DVLA sent out letters to the accused members in which they declared the recipient as being unfit to drive and in some cases order them to return their license or take a drug test (Amnesty International, 2018; Couchman, 2019; Scott, 2018). Even more concerning is that the

Gang Matrix did not clearly differentiate between potential offenders or victims, and therefore, causing unfair real-life consequences (Mayor of London, 2018).

### 5.3.2. Automated Facial Recognition Systems

Automated facial recognition (AFR) systems can identify individuals in public like at football matches, protests, or just on the street in their neighborhood (Ferris, 2018). Cameras that can be either mobile or fixed capture people's images or videos, extract facial prints, and then cross-reference with certain databases (Gates 2011). Pattern recognition techniques are widely being used, however, in the real-world environment, many challenges emerge due to acquisition condition, facial expressions, and pose alterations (Olszewska, 2016).

In Europe, automated facial recognition systems and even less Live Facial Recognition (LFR) systems are not yet rife, however, research and development in this area are done. In London, testing is conducted for public order policing like for instance the installation of facial recognition systems during the Notting Hill Carnival to identify persons of interest (Ferris, 2018; London Policing Ethics Panel, 2018). In Germany, an AFR System was used during the G20 summit in Hamburg which is still used until today. Originally, it was planned to execute reactive recognition of the violent protesters at the summit. Later, facial prints taken were cross matched against pictures from the German Federal Criminal Police Office information systems BKA INPOL police and successfully increased the chance of detecting crime (Monroy, 2018). In the Netherlands, a software called CATCH is used which cross matches images from automated teller machines (ATMs), security cameras, and social media among others against a database storing images of criminals. In Europe, no real-time facial recognition systems are deployed, however, they are developed to be implemented in the future (Jansen, 2018).

In the U.S. city Los Angeles, faces are stored and cross-matched in real-time against so-called "hot lists" including people supposedly included in gang activity or facing an open arrest warrant (Garvie & Frankle, 2016). Also, the International Criminal Police Organization (INTERPOL) developed a face recognition system that is a biometric software capable of comparing patterns in facial features and contours. The software was used first in 2016 and made approximately 60 potential matches in the following year. The INTERPOL facial recognition system automatically encodes images and produces a list of the most likely matches. Thereafter, officers have manually examined the potential candidates and act accordingly (INTERPOL, n.d.).

Several scholars proved (Buolamwini & Gebru, 2018; Hao, 2019b) that facial recognition systems perform substantially worse for black and Hispanic people, women, and other genders. As the systems are primarily trained with data of white males the system has disparities in its accuracy rates. The difference in accuracy rate can lead to misclassification of African Americans and other minorities resulting in false-positive classification making innocents wrongly remarked as suspects in crimes or even wrongly convicting them (Garvie & Frankle, 2016). Companies market their facial recognition software as highly efficient and accurate, however, in reality, these systems are not required to undergo public and independent testing to verify their accuracy or impartiality. Research claims that these communicated accuracy rates are not distributed equally and still vary in terms of race, gender, and other variables.

The “other-race effect” agrees with this finding that Western algorithms have higher accuracy rates for Caucasian faces than for East Asian faces while in contrast, East Asian algorithms achieve higher accuracy rates for East Asian faces than Caucasian faces. The accuracy advantage is, however, substantially larger for the Caucasian face and shows a generally more stable accuracy performance (Phillips, Jiang, Narvekar, Ayyad, & O'Toole, 2011). This is in line with other scholars (Garvie & Frankle, 2016; Phillips et al., 2011) who claim that algorithms from China, South Korea, and Japan better recognized East Asian faces than white people, while algorithms developed in the United States, France, and Germany have higher accuracy rates for white people. These findings indicate that the performance of the algorithms towards certain races is highly dependent on the image database used or even the development team (Garvie & Frankle, 2016; Phillips et al., 2011). As especially minority groups are being policed and investigated; therefore, are the ones relying the most on the accuracy of the recognition system, it is untenable that they are the ones suffering from the lower accuracy rates.

### 5.3.3. Sentence Length

When offenders have been convicted of a crime, AI systems still play a role in the determination of a respective penalty. Judges can inquire about criminal risk assessment algorithms for a recidivism score. This score illustrates the likelihood of a convicted criminal to re-offend in the future. Among other things, this score has an impact on the judge's decisions as if the defendant should be sent to a rehabilitation center or held in prison (Hao, 2019a, 2019b). However, models learn from statistical correlations rather than causations, meaning that if a person has similar characteristics to reoffending, the system will attribute him or her high risk. Causations which are the underlying reason are oftentimes neglected by the system. Say hypothetically, if re-offenders are oftentimes Hispanic, the system would classify according to the correlation and Hispanic people as riskier. However, skin color or origin is rather not the cause of why people would commit another crime. Still, this is how the assessment tools operate: correlations are construed as causal relationships (Hao, 2019a).

## 5.4. HEALTHCARE

AI research is becoming more and more relevant in medicine as ML algorithms are applied in complex problems (Challen et al., 2019). These algorithms can take a large amount of patient data in, able to learn unsuspected associations and thereby oftentimes outperform professionals’ estimations (Dreyfus & Dreyfus, 1992). AI definitely has the potential to disrupt the health industry, however, its entire clinical capabilities have not been exploited (Hall & Pesenti, 2017; Rao & Verweij, 2017). The restraint is mostly caused by a lack of understanding of how to guarantee patient safety while quantifying the technology’s advantages. Especially, ethical and legal concerns have hindered AI to prosper in the healthcare sector (Char, Shah, & Magnus, 2018).

Clinical decision support systems have been widely used in medicine when it comes to risk screening (Hippisley-Cox et al., 2008), prognostic scoring (Bouch & Thompson, 2008), guidelines adherence (Challen et al., 2019), and sample screening (Hippisley-Cox et al., 2008). Predefined rules decrease the risks for human clinical errors and address the issue of diagnostic uncertainty (Challen et al., 2019; Koppel et al., 2005; Nurek, Kostopoulou, Delaney, & Esmail, 2015). However, these clinical rule-based decision support systems are less suited for the breadth and variety of information needed for an optimal diagnostic.

Such a gap can be filled with diagnostic support from machine learning systems that consider a multitude of complex factors (Challen et al., 2019); thus, predicting specific diseases and personalized treatments. In the long term, the development goes, however, beyond supervised learning into the direction of proactive intervention, active learning, and reward-driven models: reinforcement machine learning. According to Challen et al. (2019), the performance of these applications is highly dependent on the quality and quantity of the training data as well as parameter selection.

Indeed, research (Chen, Johansson, & Sontag, 2018) states that prediction in mortality rates in cancer research varies in its accuracy as about 20% between the majority and protected group members. Again, this gap can be mainly explained by the sensitivity of algorithms to the quality of the training data. Another case illustrating AI discrimination in health care systems is the melanoma detection presented by Esteva et al. (2017). Convolutional neural networks have been trained to identify melanoma on images. Generally, the accuracy rates are as high as decisions made by professionals, however, influenced by various skin characteristics. These characteristics include skin type, color, and amount of hair, obviously correlating with different races and influencing accuracy rates immensely. Biased clinical trials have been documented by various scholars (Boulamwini & Gebru, 2018; Melloni et al., 2010; Popejoy and Fullerton, 2016) and have been proven to result in treatments not working sufficiently for various subgroups of the population.

# 6. SOURCES FOR BIASED AI

In order to analyze the causes for AI bias, it is expedient to look at stakeholders and steps in the model creation process. Figure 1 shows the adapted process diagram of the Cross-industry standard process for data mining (CRISP-DM) model. This process model is widely used in the Data Science community and originally defines six steps: the business understanding, data understanding, data preparation, modelling, evaluation, and deployment. In this diagram, however, a six step - the usage - is added as the way the model is used for real problem solving is a possible source of AI bias. The main stakeholders involved in the process are one or multiple coders and a user which can be a business user or any other user utilizing the developed AI system. Further, the data and the step modelling are illustrated in orange as those can also be the source of potential bias.

**Figure 1**

*CRISP-DM Model (adapted from Wirth & Hipp, 2000)*

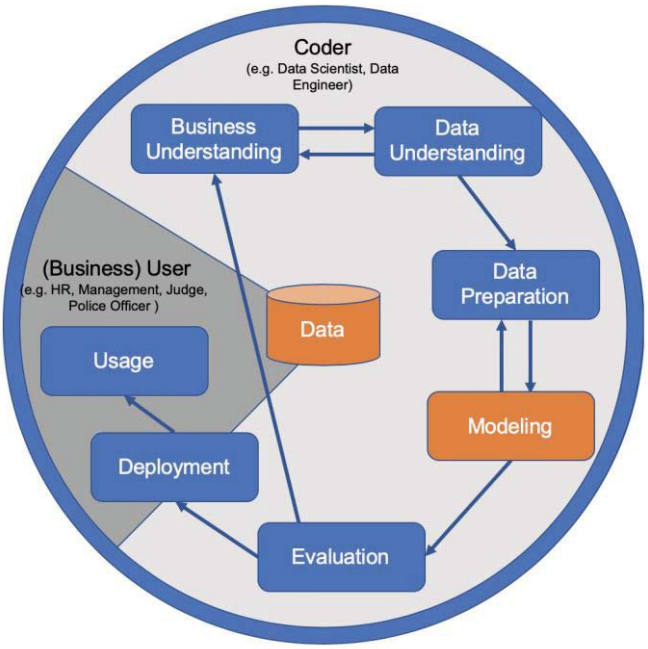
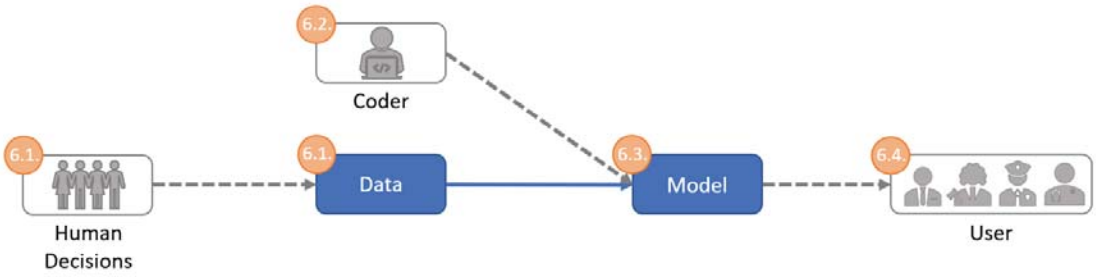



Figure 2 shows the main sources explored in this thesis for biases in AI. An AI system is based on a model that learns from past data which is based on past human decisions. This procedure is conducted by one or multiple coders who can be a Data Scientist, Data Engineer, Data Analyst, or AI Researcher. Once the model is created, it is utilized by a user in various contexts. In the following section we explore each of these sources individually.

**Figure 2**

*Sources for Biased AI*

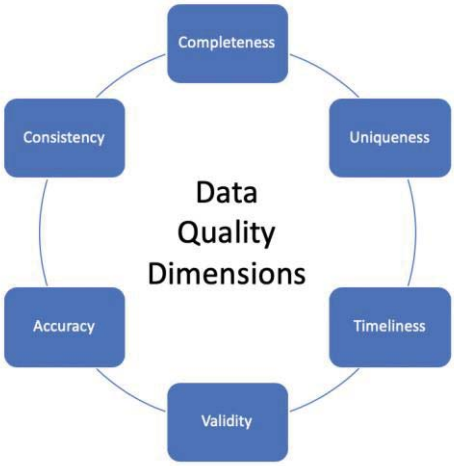



## 6.1. DATA

As data is the foundation of every model and data quality is one of the most prominent factors to ensure the value and accuracy of the models. In fact, research shows that it is essential to avoid having misrepresentations, existing human biases or incomplete knowledge incorporated in the data (Fridensköld, 2019). To evaluate the degree of data quality, a six dimensions framework can be applied. These six dimensions of data quality are: completeness, uniqueness, timeliness, validity, accuracy, and consistency (Figure 3).

**Figure 3**

*Six Dimensions of Data Quality (Askham et al., 2013)*




Completeness is measured as a proportion of the data quantity against the potential of 100% completeness. Uniqueness is achieved when each data point is only recorded once with respect to the identification. Datasets can be understood as being valid when conforming with syntax standards like format, type, and range. In addition, timeliness considers the extent to which data presents reality from the respective point in time. Accuracy describes the degree to which data constitutes the reality and respective situation while consistency is achievable when data sets and records match across data stores and datasets (Askham et al, 2013). In this section, reasons for biased data and its consequences are examined and put in relation to the framework of data quality (see Table 1).

**Table 1**

*Sources of bias in respect to the quality dimensions (adapted from Askham et al., 2013)*

| | Completeness | Uniqueness | Timeliness | Validity | Accuracy | Consistency |
|---|---|---|---|---|---|---|
| Human Bias | **X** | | **X** | **X** | **X** | **X** |
| Skewed Data | **X** | | | | | |
| Subset Targeting | **X** | | | | | |
| Outdated Data | | | **X** | | | |

### 6.1.1. Human Bias

More than 180 human biases have been identified that affect how humans evaluate, decide, and act (IBM Corporation, n.d.-a). People tend to not use sophisticated information-processing strategies, but rather simple intuitive strategies that can create biases and errors in judgment (Tversky & Kahnemann, 1973, 1974; Kahneman & Tversky, 1973, Nisbett & Ross, 1980). The Cambridge Dictionary (n.d.-a) defines bias as the "action of supporting or opposing a particular person or thing in an unfair way, because of allowing personal opinions to influence your judgment". Generally speaking, human biases result in subjective and unfair consequences which can then be encoded in datasets.

Bias in human inference can be motivational bias and cognitive bias (Kruglanski & Ajzen, 1983). A motivational bias is thereby the human tendency to form beliefs in a way to fulfill the individual's wishes and needs (McGuire, 1960). Scholars have argued that humans are motivated with regard to self-enhancement and self-protection of their egos. Concepts like "false consensus" (Ross, Greene, & House, 1977) "egocentric attribution" (Heider, 1958) can thereby help to understand how individuals tend to see their own behavior and judgments as correct as they tend to ignore or disregard alternatives views. In other words, humans are prone to biases as they are not clearly rational but incorporate self-serving perceptions in their decisions.

Even if no motivational bias existed, human biases can be originated due to the way our cognition works. This bias lies in the assumption that humans are unable to rationally and properly process all information available. One reason is sampling bias meaning that a person draws a conclusion typically from a few samples of their experiences without considering that this sample is only a fraction of the population (Kruglanski & Ajzen, 1983; Nisbett and Ross, 1980). Furthermore, Kruglanski and Ajzen (1983) state the concepts of selective attention and selective recall. To put it simply, humans access and recall information that is perceptually salient and easily available.

As data capture the past behavior, decisions, and cognitive and motivational processes of humans, the obtained data incorporates the human biases. As AI algorithms are usually trained from past

data, they naturally preserve the biases and, if not corrected, originate output highly influenced by human biases (Osoba & Welser, 2017).

In recent years, several scholars verified that an algorithm trained with biased data does create algorithmic discrimination (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016; Caliskan, Bryson, & Narayanan, 2017). The "Garbage In, Garbage Out" principle applies whereby the systems are only as good as the data provided by humans. Computers will unquestioningly process contradictory and senseless input data which will then result in nonsensical outputs (O'Neil, 2016).

### 6.1.2. Accuracy Parity

Sample size disparity can be understood as a discrepancy in the representation of groups in the training data. The reason is skewed data, meaning the data deviates from a normal distribution either to the left or right. Especially minorities are affected as they are less represented in data samples (Beutel, Chen, Zhao, & Chi, 2017).

As the model is trained with more training data from the majority than from the minority groups, the result is unfairness in Machine Learning. If 99% of the data is from the majority group, the learning algorithm will most likely not perform better than the 99% accuracy achieved by a simple classifier labeling everything with the majority class (Provost, 2000). Generally speaking, an AI rule of thumb is that the more data available, the better the model becomes. Models that learn with just limited samples are more prone to be affected by outliers and noise, and can more easily suffer from overfitting (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). Besides, certain model techniques require to train and test the model on subsets of data whereby large amounts of data is required (Srivastava et al., 2014). As the model increases its accuracy in predicting majority outputs, minority groups can get neglected as the model was not trained well enough on those types of data.

Chen et al. (2018) discussed the relation between accuracy and discrimination in their research, which has previously received relatively little attention. Providing the same predictive accuracy rates for minority groups is essential to avoid discrimination. They argue that solely the collection of additional samples can enhance fairness substantially. In contrast, widely known post-hoc methods used to defeat discrimination do so at the expense of predictive accuracy.

Research by Beutel et al. (2017) shows that even using just 500 observations of minority groups increases fairness metrics significantly. Furthermore, their experiment reveals that a more balanced dataset distribution stabilizes the model, results in smaller standard deviations, and advances the fairness of the models across all used metrics.

Especially automated facial image analysis tends to have much lower accuracy rates for all kinds of minority groups. Automated facial image analysis included various tasks such as face detection (Bai & Ghanem, 2017; Mathias, Benenson, Pedersoli, & Van Gool, 2014; Zafeiriou, Zhang, & Zhang, 2015), face recognition (Parkhi et al., 2015; Wen, Zhang, Li, & Qiao, 2016), and face classification (Buolamwini & Gebru, 2018; Levi and Hassner, 2015; Rothe, Timofte, & Van Gool, 2018). The topic received much attention when Google's automatic image labeling tool misclassified black people as gorillas (Zhang, 2015). Research from the Massachusetts Institute of Technology Media Lab (Buolamwini & Gebru, 2018) shows that leading facial recognition algorithms from IBM, Microsoft, and Face++ misclassified men below 1% at the time. However, the darker their skin color, the more

misclassification errors arise. In contrast, they misclassified up to 12% of dark men. Even worse the numbers are for females: around 7% of white females were misclassified and the worst classified group was darker-skinned females, with approximately 34,4% misclassifications. Error rates like these can mostly be explained with the data set the algorithms are trained with (Buolamwini & Gebru, 2018; Lohr, 2018). In fact, research (Han & Jain, 2014) estimates frequently used datasets are composed of 77.5% male and 83.5% white faces.

Underrepresented groups can suffer from lower accuracy rates as in law enforcement face recognition networks. At least 117 million Americans are included in these networks and investigation across 100 police departments revealed that especially Africa-Americans are more likely to be included in face recognition searches as they are more suspected and investigated (Garvie, Bedoya, & Frankle, 2016). Among all, the ones dependent on good accuracy rates are the ones facing the highest risk of misclassification. False-positive classification, and therefore, unwarranted searches or wrong convictions are threatening and discriminating against minority groups. Especially women, black people, and young people are under the risk of being misclassified (Klare, Burge, Klontz, Bruegge, & Jain, 2012).

### 6.1.3. Subset Targeting & Simpson's Paradox

Edward Hugh Simpson denotes a statistical phenomenon in 1951 which takes his name "Simpson's Paradox". This paradox describes the case when combined outcomes show a different pattern from when certain subgroups are analyzed individually (Pearl, 2014). In 1973, one of the most prominent Simpson's paradox cases occurred at the University of California, Berkeley when they were accused of being gender biased. From those applying, 44 percent of male graduates and 25 percent of female graduates were admitted. The perception flipped, however, when the data was viewed from another angle. When broken down by departments, the acceptance rate of females was actually higher in 4 of the 6 departments and no substantial difference could be found in the other two. A closer analysis showed the force of the Simpsons' Paradox; females applied to faculties that generally accepted a smaller percentage of applicants, and therefore, reversing the overall distribution (Grigg, 2019).

Scholars (Dwork et al., 2012) related Simpson's paradox to fairness and revealed that distinct protective subgroups can be hidden within a larger subgroup. This larger subgroup can be defined as statistically neutral and can, therefore, cover potential discrimination. To put it in other words, an individual can be categorized as black whereby he is, according to statistical measures, not discriminated against. However, several different origins can be identified as black and might be statistically not discriminated against. The high number of negative outcomes of this smaller subgroup can then be balanced out by the overall more averaged amount of the larger subgroup. Oftentimes, no awareness of this kind of discrimination exists as the smaller subgroups are not identified and more generic group definitions are used in datasets and hence, biases cannot be identified not can measures be taken.

### 6.1.4. Outdated Data

Biased AI can be concluded from an outdated perception of our society which is encoded in outdated data. Therefore, it is essential to analyze if data fed to the model is still contemporary at the moment the model is built and at the moment the model is being used.

## 6.2. CODER

When analyzing AI algorithms, it is essential to examine the creators of the algorithm. Findings from the AI Now Institute team, at New York University, discovered that 82 percent of the authors attending leading AI conferences are male and that similarly, 80 percent of AI professors are men. In the business world, similar numbers can be found: 85% of Facebook and 90% of Google's AI research staff are men (Houser, 2019). The lead of the ethical Artificial Intelligence Team at Google Timnit Gebru shared her experience of being one of the only six black people attending a ML conference with 8.500 attendees in 2016. At Google solely 2.5% of their employees are black, while Facebook and Microsoft comprise only 4% black employees.

Furthermore, the people who create powerful AI algorithms belong to a certain income group way above the average citizen. The New York Times mentioned that AI specialists without any or just little industry experiences earn between $300,000 and $500,000 per year in wages and stocks. Top AI researchers with extensive knowledge receive salaries extending into millions. For instance, at the London AI lab DeepMind that was bought by Google, the costs for their 400 employees sum up to $138 million in 2016 translating into an average of $345,000 per employee (Metz, 2018).

Stats like these illustrate the lack of gender and racial diversity and show that primary white rich men are in charge of creating these powerful algorithms. Stanford researcher Danaë Metaxa has a similar perception stating "The urgency behind this issue is increasing as AI becomes increasingly integrated into society … Essentially, the lack of diversity in AI is concentrating an increasingly large amount of power and capital in the hands of a select subset of people." (as cited in Paul, 2019, para. 5). As mostly one homogenous group creates the models making decisions about all kinds of people of different races and colors, different social classes and genders, it is questionable if they have their best attention at heart. To go one step back, they might not have bad intentions, however, not the consideration or awareness of others not alike people.

The AI Now Institute, an interdisciplinary research center that focuses on the social implications of AI, recommends in their research paper to employ more people of color, females and other minority groups at leadership positions in AI companies (AI Now Institute, n.d.; West, et al., 2019). No scientific proof in research can be found that the homogeneity of AI professionals has negative implications on biased AI. However, there are reasons to believe that a more diverse workforce can result in more attention and focus to create fair and unbiased systems.

## 6.3. MODEL

While ML can be understood as data being parsed through algorithms to get an outcome, an ML model or AI model can be described as a relation between variables based on certain algorithms and other settings (Wong, 2018). Models can be for instance in the form of a simpler Linear Regression Model, Decision Trees, Logistic Regression, K-Nearest Neighbors, or a more complex Neural Network Model, just to name a few. In what manner AI models can contribute to biases because of certain bugs, misconstructions and other flaws will be discussed in the following section.

### 6.3.1. Feedback Loops

O'Neil (2016) illustrates in her book the situation of the criminal justice system and its problem which will be defined as a feedback loop in this paper. Especially black people are targeted and

discriminated against by the algorithms used in the criminal defense system in the United States. AI systems are being used for identifying potential criminals which especially target black men. Indeed, the reoffending likelihood prediction tool used in the United States falsely predicts a higher risk of recidivism for black defendants than they actually pose (Cossins, 2018). The reason for this is that if, for instance, a neural network is trained with biased data, the algorithm will replicate the bias. Consequently, the biased outcome will result in biased decisions that are then used as new input data for the model. Decisions made by the algorithm will become slightly more unequal with additional decisions made and inputs created. The result is a problematic feedback loop that gradually becomes more unequal and biased, creates more and more unfairness with which outcomes we actually train, and supposedly improve our models (Casacuberta, 2018).

Predictive analytic tools like PredPol, HunchLab, and CompStat are used in budget-strapping police departments around the United States (O'Neil, 2016) to forecast potential crime. PredPol takes old crime data like type, location and time and associates it with socio-economic data to predict crime happening in the next 12 hours (O'Neil, 2016; Smith, 2018). As the software targets geography and not individuals, PredPol founder and UCLA professor Jeffrey Brantingham (as cited in O'Neil, 2016) claims the model he built does not discriminate against race and ethnicity. However, even if the algorithm claims to be colorblind, the results are not (O'Neil, 2016). As geography and poverty are highly correlated with race, the system arguably will indirectly target certain social groups more often than others. The system then categorizes these black and Hispanic areas as "high risk" and consequently officers are requested to spend a minimum of 10% of their patrol time there. Even if the crime rate of districts with more dark-skinned citizens does not differ from the rate of the white neighborhoods, patrolling cops in these areas increases, so does the likelihood of the detection of any crime. A feedback loop is created whereby the association of black districts with high risks enforces higher crime rates, which reinforces the system to gain more focus on these areas. The combination of the redlining effect and the feedback loop leads to unintended discrimination which is still unfair to the dark-skinned society. Even if the tool has its best intention to decrease crimes, it targets the poor, searches them, arrests a fraction of them and even a subgroup is sent to prison. Just in a statistical manner, the dark-skinned people are disadvantaged as even if crime rates are the same, the likelihood of being arrested increases because of these tools.

The reinforcement loop can be understood as a continuous circle whereby the algorithm reinforces itself. The data used in the beginning to train the model will most likely shape the model for a lifetime. Once the model detects a direction it will follow this way and it is arduous to change its embossing. This phenomenon can be easily illustrated with the Netflix recommender system. For instance, a customer who has selected a few romance movies will be shown suggestions for other movies similar to his previous preferences and will increase the likelihood of the customer watching more movies of this kind. Once this person takes the path, the system will get positive feedback and will be reassured to continue with related movie suggestions. The more the system is reinforced, the less likely it is to show other kinds of movies and thus, the customer is less likely to explore other genres. The same concept, but with more severe consequences, can be seen in the world of bias AI resulting in discrimination and unfairness. The crime detection systems explained above follow the same course; the system learns based on old data predicting that in black neighborhoods crimes are most likely to happen. Once these predictions are made, the police start patrolling, arrest people from these districts and the system creates new data pointing a finger on these districts again.

In contrast, some models might not get any feedback at all, hence the model has no chance to improve. In her book, O'Neil (2016) describes the misfortune of a fifth-grade teacher being fired based on a so-called IMPACT score. A tool was applied in the Washington D.C. schools to improve the local underperforming schools. Teachers performing below the threshold were fired. The model had no chance of evaluating its performance nor improving. This phenomenon can be also found beyond firing such as the invitation process for job interviews or university admission. When candidates get rejected, no evaluation of their potential performance can be collected, hence, no feedback on potential wrongdoing can be included in the model. The system creates its own reality and ensures continuation due to its own reinforcement.

### 6.3.2. Black Box Algorithms

Incrementally, predictive models contain millions of parameters, therefore, leveraging variable relations not noticeable nor comprehensible to humans. This transition has caused greater efficiency, productivity, and accuracy; however, it comes with undesirable fallouts. With the growing number of variables, the model's complexity increases as well. Consequently, models have been widely accused of being black boxes. These black box models can often generate highly accurate outcomes as they consider multiple different factors. However, an in-depth understanding of the derivation of the result is not given. Thus, practitioners import data into models and produce results without having the comprehension of why their models are functioning well. In areas such as speech recognition or computer visions, whereby signal structures are recognized, a lack of understanding can be excused, states Adebayo and Kagal (2016). However, in fields like banking, insurance, and employment where the accuracy is essential for livelihood, it is more important for practitioners applying the models to truly and deeply comprehend the internal workings behind it.

Reinforcement machine learning algorithms, especially those based on artificial neural networks tend to be more complex, and therefore, less comprehensible (Challen et al., 2019). When hidden layers are being used, algorithms take on a form of a more intricate construct which are oftentimes not even for its creator understandable (Alain and Bengio, 2016; Schwartz-Ziv & Tishby, 2017). For instance, Google applied a 22-layer deep model called GoogLeNet model with an amount of overall 200 layers considering the independent building blocks (Szegedy et al., 2015). Models like these, clearly are not understandable either to the average people or to coders. As the black box models make inscrutable decisions, it is hard to detect or even to be aware of errors and biases of the model. Challen et al. (2019) illustrate this issue with an AI experiment, whereby the U.S. army used ML to differentiate between pictures of trees obscuring armored vehicles and just plain forests. The system was performing equally to a guess probability, since the images the system was trained with had been taken inclusively in the sun and did not match the cloudy conditions in the field. Issues like these can lead to severe, in this case even life-threatening, consequences. As the systems are highly complex, the average army officer has no chance of knowing the source of the problem.

Besides, people that are not actively using predictive models can still be affected by the measures. O'Neil (2016) describes several occasions whereby normal citizens fall victim to AI tools. For instance, people being fired, not admitted to a university, not granted a loan, or not being offered an apartment, just to name a few. These occurrences have in common a negative outcome for an individual without a proper explanation or well-grounded information. Obtaining this information requires usually extraordinary effort closely linked to monetary costs. Oftentimes getting an attorney

is required resulting in expensive billings that have to be evaluated against the not assured benefits. Even if monetary resources are not an issue, those concerned are oftentimes not even aware of the negative outcome being provoked by a biased algorithm. However, even if the legal representation is successful to get the algorithm's information, the next step to analyze and understand the model's defects is another hurdle for almost anyone. Once the victim gets the algorithm, the task of opening the black box model most likely remains.

### 6.3.3. To-Be vs. Should-Be Model

The general concept of how models are created causes reason to be rethought. Models use past data and assume and reassure that the future is shaped identically to the past. To simplify, imagine a recruiting tool deciding who gets invited to an interview. Models perceive past decisions as optimal choices and invite candidates based on similar profiles. However, they do not provide space for new diversified profiles to be interviewed. When no varying profiles receive interview invitations, their performance in the interview or potentially in the job cannot be evaluated. How can the model then know that it was good to not invite these people in the first place? It cannot. This is the fatal consequence of believing models can solely translate past behavior into prospective optimal actions. It is believed that by taking the insights of an AS-IS model, a TO-BE model can be created which depicts our current needs, perceptions, and ethics. However, a platform for discrimination, bias, and unfairness is fabricated providing no space for diversification and change.

To fill this gap, this paper introduces the concept of a Should-Be model. It should be a thought-provoking impulse that reconsiders the whole model construction. Rather than including old perceptions, biases, and habits, a should-be model should factor in contemporary standards, ideals, and future demands to discontinue passing on old biases. For this purpose, not only developers should be involved in the model creation process, but also people with an ethical and social understanding to not only perpetuate old standards and transfer outdated and unfair decisions into the future, but rather developed models that create fair and social decisions.

### 6.3.4. Correlation vs. Causation

Many machine learning methods use correlations to make predictions. Generally, correlations denote that certain phenomena occur together and are used due to their potential predictive power. One value is used to predict another correlating value. Already in the 10th century BC, early astronomers in Korea and China computed future observation of astronomical bodies from multiple regularities. In the 10th century in China, doctors related to the infection of mild smallpox with the prevention of severe illnesses. An experiment conducted by Needham (1976) successfully confirms this correlation. Correlations have demonstrated its practical relevance; however, one should be aware of its limitations. In the real world, mathematical correlations do not need to translate to causations. In the 1940s, children were advised to not eat ice cream due to a correlation of Polio infections and eating ice cream (Flovik, 2019). Medically no relation between the illness and the sweet exists, the only relation was that Polio epidemics occurred during summer and fall where children were eating more ice cream. This misconception illustrates the confusion between correlation and causation. In this instance, the two occurrences were not correlating with each other but with a third factor: the time of the year. The same misconceptions can be found in the direction of the correlation. One thing can cause the other but does not have to be a mutual relationship.

The risk to confuse correlation and causation remains also in machine learning since algorithms cannot distinguish between correlation and causation (Leetaru, 2019). For instance, a correlation between race and university admission can exist, however, it does not guarantee causation indicating a specific race is better suited for a university place.

As an abundance of data is generally related to a higher accuracy rate, this inference should be carefully analyzed. Scholars conclude that as the amount of data increases, algorithms will find correlations and regulation no matter the meaning or content of these correlations. Anderson (as cited in Calude & Longo, 2017, p.17) stated that ‘‘Correlation is enough (....) We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot’’. The more data, the more meaningless correlations can be found. Paradoxically, in this sense, more data results in increased difficulty to differentiate between noise and insights.

### 6.4. User and Model Usage

Model deployment is another crucial source of potential biases. As AI models produce results which most often are not automatically deployed, a human is needed to carry out the model’s decision. This means that there is a last instance which has the final power over the decision. For instance, an AI tool can predict in which district crime is most likely to occur. However, a police officer is the one who is executing the model's decision. Same goes for an AI tool predicting the likelihood of a criminal to reoffend. Here, a judge is the one who makes the final ruling. In other sectors, like the employment sector where an AI system is scanning CV and decides who to invite for an interview, the process might be automated or also involves a human recruiter who has the final say.

Anyhow, it is important to emphasize that AI models are not always flawless and that biases can be incorporated. As definitions and statistical measures of fairness can establish fairness, they cannot evaluate the social context in which the AI system is used (Silberg & Manyika, 2019). It is therefore important to consider where and in what form human judgment is necessary (Silberg & Manyika, 2019). Therefore, not only technical professionals included in the model creation process must be educated on potential risks and AI ethics, but also people in all kinds of areas using the models have to be informed. Therefore, it is important to critically question AI outcomes and understand results rather as a suggestion than call to action. Research has shown the advantages of AI models and the capability with which they can perform more complex decisions than humans. However, trusting results without any verification, comes with risk.

# 7. SOLUTIONS FOR BIASED AI

## 7.1. FAIRNESS DEFINITIONS & METRICS

In order to potentially mitigate biases in AI systems, tools to measure bias and fairness have to be provided and critically evaluated. Therefore, Table 2 was created to provide an overview and cluster the abundance of measurements available in research. Verma and Rubin (2018) describe in their paper "Fairness Definitions Explained" various fairness measures that can be used in the field of algorithmic fairness which is extended with additional research and compiled in a tabular format in this thesis.

Verma and Rubin (2018) differentiate between statistical metrics (e.g. definitions based on the statistical outcome, definitions based on predicted and actual outcomes, measures based on predicted probabilities and the actual outcome), similarity-based measures, and causal reasoning (see Figure 5). In addition, Zehlike, Castillo, and Bonchi (2017) provided further measures grouped into absolute measures and statistical tests. All these measures solely focus on providing fairness measures for supervised algorithms.

**Table 2**

*Overview measures biased AI (adapted from Verma and Rubin, 2018; Zehlike et al., 2018)*

| Measure | Explanation/ Formula | Source |
|---|---|---|
| **General Statistical Metrics (confusion matrix, accuracy)** | | |
| **Positive Predictive Value (PPV)** | The proportion of the truly predicted positives to all truly and falsely predicted positives.<br>PPV = TPTP + FP | E.g. Verma & Rubin, 2018 |
| **False Discovery Rate (FDR)** | The proportion of falsely predicted positives to all truly and falsely predicted positives.<br>FDR =FPTP + FP | E.g. Verma & Rubin, 2018 |
| **False Omission Rate (FOR)** | The proportion of the falsely negative predicted values to all falsely and truly predicted negatives.<br>FOR = FNTN + FN | E.g. Verma & Rubin, 2018 |
| **Negative predictive value (NPV)** | The proportion of the truly predicted negatives to the true negatives and falsely predicted negatives.<br>NPV = TNTN+FN | E.g. Verma & Rubin, 2018 |
| **True Positive Rate (TPR)** | The proportion of the truly predicted positives to all the actual positives.<br>TPR = TPTP + FN | E.g. Verma & Rubin, 2018 |
| **False Positive Rate (FPR)** | The proportion of the falsely predicted positives to all the actual negatives.<br>FPR = FPFP + TN | E.g. Verma & Rubin, 2018 |
| **False Negative Rate (FNR)** | The proportion of all the falsely predicted negatives to all the actual positives.<br>FNR = FNTP + FN | E.g. Verma & Rubin, 2018 |

| | | |
|---|---|---|
| **True Negative Rate (TNR)** | The proportion of all truly predicted negatives to all the actual negatives. TNR = TNFP + TN | E.g. Verma & Rubin, 2018 |
| **Definitions Based on Predicted and Actual Outcomes** | | |
| **Predictive parity [a], outcome test [b]** | Same PPV for protected and unprotected groups. | [a] Chouldechova (2017) [b] Simoiu et al. (2017) |
| **False positive error rate balance [a], predictive equality [b]** | Same FPR for protected and unprotected groups. | [a] Chouldechova (2017) [b] Corbett-Davies et al. (2017) |
| **False negative error rate balance [a], equal opportunity [b]** | Same FNR for all protected and unprotected groups. | [a] Chouldechova (2017) [b] Hardt, Price, & Srebro, (2016); Kleinberg, Mullainathan & Raghavan (2016) |
| **Equalized odds [a] (a.k.a. conditional procedure accuracy equality [b] and disparate mistreatment [c]** | Same FPR and same TPR for protected and unprotected groups. | [a] Hardt et al. (2016) [b]Berk, Heidari, Jabbari, Kearns, & Roth (2018) [c] Zliobaite (2015) |
| **Conditional use accuracy equality** | Same PPV and same NPV for protected and unprotected groups. | Berk et al. (2018) |
| **Overall accuracy equality** | Same prediction accuracy for protected and unprotected groups. | Berk et al. (2018) |
| **Treatment equality** | Same ration of false negatives and false positives for the protected and unprotected groups. | Berk et al. (2018) |
| **Measures based on Predicted Probabilities and Actual Outcome** | | |
| **Test-fairness [a] (a.k.a. calibration [a], matching conditional frequencies [b])** | Test-fairness is achieved if any predicted probability score for the protected and unprotected class truly belonging to the positive class have the same probability score. | [a] Chouldechova (2017) [b] Hardt et al. (2016) |
| **Well-calibration** | Well-calibration is an extension of the definition of test-fairness. For each predicted probability score, individuals in both protected and unprotected classes should not merely have the identical probability having a positive outcome, further should have probability equal to the probability score. | Kleinberg et al. (2016) |

| | | |
|---|---|---|
| **Balance for positive class** | Individuals belonging to the positive class from both protected and unprotected groups have the same average predicted probability score. | Kleinberg et al. (2016) |
| **Balance for negative class** | Individuals belonging to the negative class from both protected and unprotected groups have the same average predicted probability score. | Kleinberg et al. (2016) |
| **Definitions based on statistical outcome (Group Fairness)** | | |
| **Group fairness [a] (a.k.a. statistical parity [a], equal acceptance rate [b], benchmarking [c])** | Group fairness is achieved if individuals of the protected and unprotected group have the same probability of belonging to the positive predicted class.<br><br>P(outcome = 1\| Protected Variable) = P(outome= 1\|unprotected variable ) | [a] Dwork et al. (2012)<br>[b] Zliobaite (2015)<br>[c] Simoiu, Corbett-Davies & Goel (2017) |
| **Conditional statistical parity** | Conditional statistical parity is achieved if individuals of the protected and unprotected group have the same probability of belonging to the positive class. This definition includes a set of attributes that can affect the outcome.<br><br>P(outcome = 1 \| protected variable I, protected variable II) = P(outcome = 1 \| unprotected variable I, unprotected variable II ) | Corbett-Davies, Pierson, Feller, Goel, & Huq (2017) |
| **Similarity-based measures (Individual Fairness)** | | |
| **Causal discrimination** | Causal discrimination defines if a classfier achieves the same classification for any two subjects who have exactly the same attributes but can differ in its protected attributes. | Galhotra, Brun, & Meliou (2017) |
| **Fairness through unawareness** | Fairness through awareness is achieved when only protected attributes are not explicitly included in the decision. This implies that the classification outcome is equal for all applicants with the same attributes. | Kusner et al. (2017) |
| **Fairness through awareness** | In this definition, fairness is defined as such that similar individuals are supposed to have similar outcomes. Distance metrics are used to assess the similarity of individuals and their outcomes. Fairness is achieved if the distribution of outcomes is at most the distance between the individuals. | Dwork et al. (2012) |
| **Causal Reasoning** | | |

| | | |
|---|---|---|
| **Counterfactual fairness** | Counterfactually fair if the predicted outcome does not depend on a descendant of the minority group's attribute in the causal graph (aims to address individual fairness). | Kusner et al. (2017); Salimi, Howe & Suciu (2019) |
| **No unresolved discrimination** | No unresolved discrimination if there is no connection between the potected attribute to the outcome (exept over a resolving variable) in the causal graph. | Kilbertus et al. (2017) |
| **No proxy discrimination/ Proxy Fairness** | No proxy discrimination if no connection between the protected attributed and the predicted outcome in the causal graph. | Kilbertus et al. (2017) |
| **Fair inference** | Fair inference if no illegitimate paths from the redicticed attribute to the predicted outcome. | Nabi & Shpitser (2018) |
| **Interventional Fairness** | Interventionally fair if protected attributes do not affect the predictor in any formation (aims to address group fairness). | Salimi et al. (2019); Salimi, Rodriguez, Howe & Suciu (2019) |
| **Path-specific fairness** | Path-specific fairness describes the illicit path-specific causality from connections between sensitive attributes to the outcome | Loftus, Russell, Kusner, & Silva (2018); Salimi, et al. (2019) |
| **Absolute Measures (Magnitude of Discrimination)** | | |
| **Mean Difference** | Difference between the target score mean values of the protected in comparison to the unprotected groups. No difference indicates no discrimination. | Zehlike et al. (2017) |
| **Normalized Difference** | Mean difference normalized by the rate of positive outcomes. | Zehlike et al. (2017) |
| **Impact Ratio (Zehlike et al., 2017) / Group Fairness (Verma & Rubin, 2018)** | The ratio of positive outcomes for the protected group over the non-protected group. | Zehlike et al. (2017) |
| **Odds Ratio** | The ratio between exposure and outcome | Zehlike et al. (2017) |
| **Difference of Means (Welch-Test)** | Null hypothesis that the means of the protected and non-protected groups are equal. | Zehlike et al. (2017) |
| **Difference in proportions for two groups (Fisher's exact test)** | Null hypothesis that rates of positive outcomes for the unprotected and protected groups are equal. | Zehlike et al. (2017) |

| | | |
|---|---|---|
| **Difference in proportions for many groups** | Null hypothesis that the probabilities or proportions are the same for all (protected and unprotected) groups. | Zliobaite, 2017 |
| **Regression slope test** | Null hypothesis that tests if the regression coefficient of the protected variable is substantially different from zero. | Zliobaite, 2017 |
| **Rank Test** | Null hypothesis that the distributions of the protected and unprotected groups are equal. | Zliobaite, 2017 |
| **Rank Test/ z-score (for large Groups)** | Normal approximation and then the z-test can be used to test the distribution of the groups. | Zliobaite, 2017 |
| **Difference of Means (Welch-Test)** | Null hypothesis that the means of the protected and non-protected groups are equal. | Zehlike et al. (2017) |
| **Difference in proportions for two groups (Fisher's exact test)** | Null hypothesis that rates of positive outcomes for the unprotected and protected groups are equal. | Zehlike et al. (2017) |
| **Difference in proportions for many groups** | Null hypothesis that the probabilities or proportions are the same for all (protected and unprotected) groups. | Zliobaite, 2017 |
| **Regression slope test** | Null hypothesis that tests if the regression coefficient of the protected variable is substantially different from zero. | Zliobaite, 2017 |
| **Rank Test** | Null hypothesis that the distributions of the protected and unprotected groups are equal. | Zliobaite, 2017 |
| **Rank Test/ z-score (for large Groups)** | Normal approximation and then the z-test can be used to test the distribution of the groups. | Zliobaite, 2017 |
| **Difference of Means (Welch-Test)** | Null hypothesis that the means of the protected and non-protected groups are equal. | Zehlike et al. (2017) |
| **Difference in proportions for two groups (Fisher's** | Null hypothesis that rates of positive outcomes for the unprotected and protected groups are equal. | Zehlike et al. (2017) |

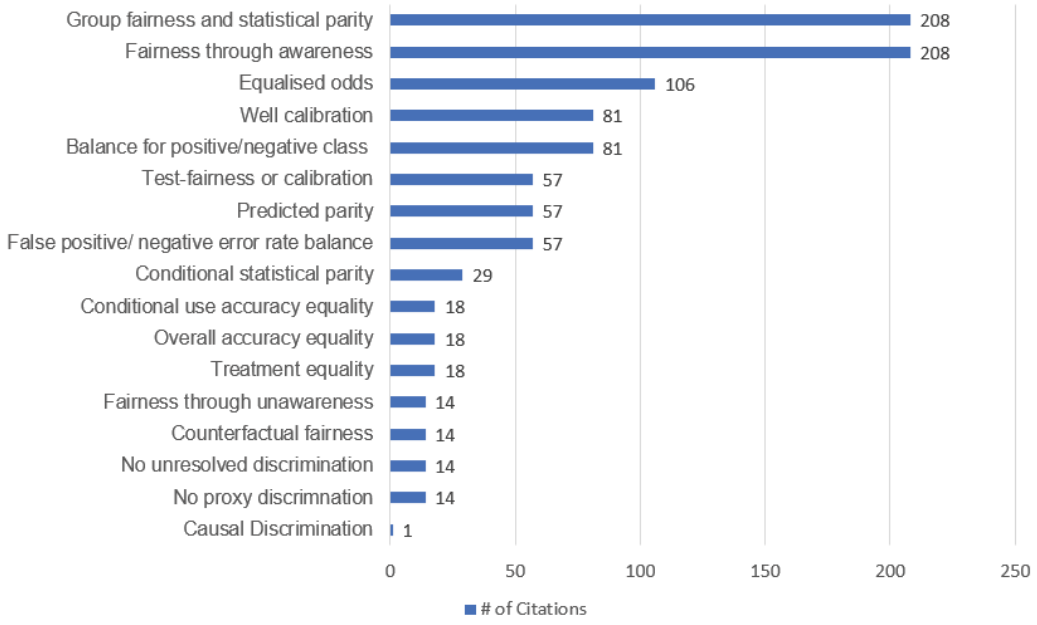
| | | |
|---|---|---|
| **exact test)** | | |
| **Difference in proportions for many groups** | Null hypothesis that the probabilities or proportions are the same for all (protected and unprotected) groups. | Zliobaite, 2017 |
| **Regression slope test** | Null hypothesis that tests if the regression coefficient of the protected variable is substantially different from zero. | Zliobaite, 2017 |

Verma and Rubin (2018) provided data on the number of citations from the different measurements which were reprocessed and visualized in Figure 4. The measurements "group fairness and statistical parity" and "fairness through awareness" have according to their recorded numbers 208 citations each. The number of citations can be understood as a favorability, presence, and importance in research.

In the following section we explore the main categories of measured presented in Table 2.

**Figure 4**

*Numbers of Citations for the Main Fairness Metrics (data from Verma and Rubin, 2018)*




### 7.1.1. General Statistical Metrics

The general statistical metrics listed in Table 2 (Verma & Rubin, 2018; see also Provost & Kohavi, 1998) form the foundation of more advanced metrics. In that sense, it is relevant to recall the underlying measures behind a confusion matrix which is displayed in Table 3. The matrix puts into relation the actual and predicted values in a classification model and is made up of four metrics: true

positive (TP), false positive (FP), false negative (FN) and true negative (TN). These metrics are widely used in the data science community to capture the accuracy of a classification model (Provost & Kohavi, 1998).

**Table 3**

*Confusion Matrix (Provost& Kohavi, 1998)*

| | **Actual Positive** | **Actual Negative** |
|---|---|---|
| **Predicted Positive** | True Positive (TP) | False Positive (FP) |
| **Predicted Negative** | False Negative (FN) | True Negative (TN) |

With the categories formed, further metrics such as the Positive Predictive Value and False Discovery Rate are created putting all the categories into different relations which are listed in Table 2.

### 7.1.2. Definitions Based on Predicted and Actual Outcomes

The definitions based on predictive and actual outcomes take not only into consideration the predicted outcome but also the actual outcome captured in the dataset (Verma & Rubin, 2018). These metrics set conditions for the minority and majority class to have equal outcomes of the statistical metrics proposed before for fairness to hold true. Predictive parity (Chouldechova, 2017), for instance, is achieved when the Positive Predictive Value (PPV) is the same for the protected and unprotected group.

### 7.1.3. Definitions Based on Statistical Outcomes

The definitions listed in the third part belong the most straightforward notions of fairness namely group fairness and conditional statistical parity. Group fairness, also called statistical parity, is achieved if individuals of the protected and unprotected group have the same probability of belonging to the positive predicted class (Dwork et al., 2012). Conditional statistical parity is achieved if individuals of the protected and unprotected group have the same probability of belonging to the positive class. This definition includes a set of attributes that can affect the outcome (Corbett-Davies, Pierson, Feller, Goel, & Huq, 2017).

These two types of group fairness are widely applied to measure fairness, however, lacks to include certain features. First, measure solely applies for a prescribed protected variable and does not consider combinations (Binns, 2020). For instance, Afro-American women can be discriminated against. The statistical parity measure will only assess discrimination for the protected group “Afro-American” and “women” separately and therefore, cannot detect unfairness.

A second problem identified for group fairness is that statistical parity does not take into consideration other variables other than the protected group variable and outcome. An individual from the minority group can belong with the same probability to the positive class, however, has to perform better in order to do so. For instance, Asia-Americans make up a proportional high stake of U.S. elite universities. As they make up only 5 percent of the students attending U.S. High Schools, they contribute roughly 18.6 percent of all undergraduate and graduate students who were enrolled

in 2019 to 2020 at Massachusetts Institute of Technology and represent 58.4% of all minority students (excluding international students)(Massachusetts Institute of Technology, 2020). At Harvard University, 22% of the new freshmen were Asian-American (Jaschik, 2017). In respect to statistical parity, Asian-Americans would not be discriminated against as High School Students are being admitted with a proportionally high stake at U.S. Ivy League schools. Still, lawsuits have been filed on behalf of Asian-Americans as they perceived discrimination since they have to perform substantially better in admission criteria such as test scores, grades and extracurricular activities than any other racial group to be admitted (Hartocollis, 2018). In 2018, a lawsuit against Harvard University was issued as they consistently lowered the Asian-American applicants score for personality traits such as positive personality, likeability and kindness amongst others (Hartocollis, 2018). This example illustrates the deficiency of statistical parity as a fairness measure since no other criteria are considered like for example the performance of the applicants.

Literature roughly defines individual fairness as the counterpart to group fairness. While group fairness is focusing on groups defined by protected characteristics receive similar treatments or results, individual fairness holds true when similar persons receive similar treatments or outcomes (Bellamy et al., 2018). Therefore, individual fairness can solve some certain problems that arise in group fairness measures which will be discussed in chapter 7.1.5.

### 7.1.4. Definitions Based on Predicted Probabilities and Actual Outcomes

In part four of Table 2, measures based on predicted probabilities and actual outcome are listed. These methods take into consideration if individuals truly belong to a certain positive group.

Test-fairness is achieved when individuals from protected and unprotected groups have the same probability score of truly belonging to the positive class. This definition has resemblances to predictive parity (see 7.1.2.), apart from considering the fragment of true positive predictions for every value of the probability score (Verma & Rubin, 2018).

Well-calibration builds on the definition of test-fairness as the protected and unprotected group should not only have the same probability of truly having a positive outcome but also the same predicted probability score (Kleinberg, Mullainathan & Raghavan, 2016; Verma & Rubin, 2018).

The definition balance for positive class is oriented towards the equal opportunity definition and is satisfied if the protected and unprotected group have the same mean predicted probability score (Kleinberg, Mullainathan & Raghavan, 2016; Verma & Rubin, 2018). The definition balance for negative class is the counterpart version and states that the negative class for both protected and unprotected groups should have the same mean predicted probability score (Kleinberg, Mullainathan & Raghavan, 2016; Verma & Rubin, 2018). Both definitions lack to consider individuals as it considers the mean individual probability score can balance each other out meaning that certain individual unfairness can remain undetected.

### 7.1.5. Similarity-based Measures

In contrast to the proposed statistical metrics (see 7.1.3.), similarity-based metrics do not only take into consideration the sensitive attribute, but also other insensitive variables. That way, discrimination will be discovered as for instance minorities might have the same positive rates, however, they might have to perform better within the other attributes for their achievement

(Galhotra, Brun, & Meliou, 2017). The discussed issue (see 7.1.3.) that Asia-Americans are allegedly discriminated against as they have to perform substantially better to be admitted to Ivy league university is therefore not disguised by the high probability rate of being admitted to universities. A statistical fairness measure will not detect discrimination and will potentially declare the model to be fair, while similarity-based measure takes into consideration insensitive attributes such as an applicant's high school grades.

Causal discrimination applies when a classifier generates an equal classification for every two individuals with the same exact attributes (Galhotra et al., 2017; Verma & Rubin, 2018). In the previous example, this would mean that an Asian-American applicant with the same attributes such as grade, test score and extracurricular activities should have the same outcome as any other American with the same attributes. This definition is efficient as it does not generalize the unprotected attributes and provides a more thorough assessment. However, when an abundance of unprotected attributes is given for a smaller data set, the comparison of individuals to a similar persona might be limited. Another issue in certain cases can be the redlining effect (see chapter 3.3.2.) as certain unprotected attributes might correlate with the protected group. A certain individual might be, thereby, only compared with certain individuals from the protected group and might have a generally lower outcome. With respect to college admission, black students could be discriminated against as they might be just compared with individuals from the same High School which happens to be in a neighborhood with a proportionally higher stake of Afro-American citizens. In this case, the definition would falsely classify this case as not being discriminating.

Kusner et al. (2017) definition of fairness through unawareness is satisfied if no sensitive variables are explicitly included in the classification process. This means that protected features such as gender, age, and race are not used in the training process. The definition can be tested when the model trained without any protected attributes achieves the same result for an individual with the same features, however, with different protected attributes (Verma & Rubin, 2018). Thereby, it is ensured that no other proxy for the protected attribute is used (Vera & Rubin, 2018) meaning that the redlining effect can be detected.

Fairness through awareness by Dwork et al. (2012) is one of the most widely used and discussed measurements in research (see Table 2). This definition aims to avoid the misconception of group fairness and in contrast targets individual's fairness. It is an elaborated and general revision of the previous discussed measures. Fairness is understood as the percept that similar individuals should have similar outcomes. Therefore, a distance metric is developed which depicts the distance between individuals. Fairness is thereby achieved if the distance between the distribution of outcomes for individuals is not exceeding the distance between these individuals (Verma & Rubin, 2018).

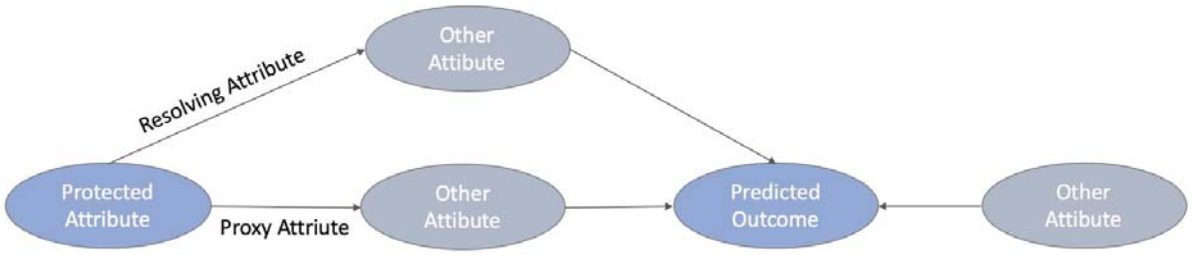
### 7.1.6. Causal Reasoning

Finally, causal reasoning assumes a specific causal graph displayed in Figure 5: a directed, acyclic graph with nodes symbolizing attributes of a requester and edges depicting relationships between the attributes. These graphs are used to create fair machine learning algorithms as they map the relationships between attributes and their impact on the result are addressed by a set of structural equations. These equations are further utilized to develop methods for assessing the impact of sensitive attributes and for building algorithms that guarantee an acceptable degree of

discrimination based on the respective attributes (Kilbertus et al., 2017; Kusner, Loftus, Russell, & Silva, 2017; Nabi & Shpitser, 2018). Figure 5 illustrates a general form of a causal graph and shows how certain protected and unprotected attributes influence the predicted outcome. A proxy attribute is defined as an attribute from which the protected attribute can be concluded. A resolving attribute can be understood as an attribute that is affected by the protected attribute, however, not in a discriminatory form.

**Figure 5**

*General Form of a Causal Graph (adapted from Verma and Rubin, 2018)*




## 7.2. Bias Mitigation Techniques

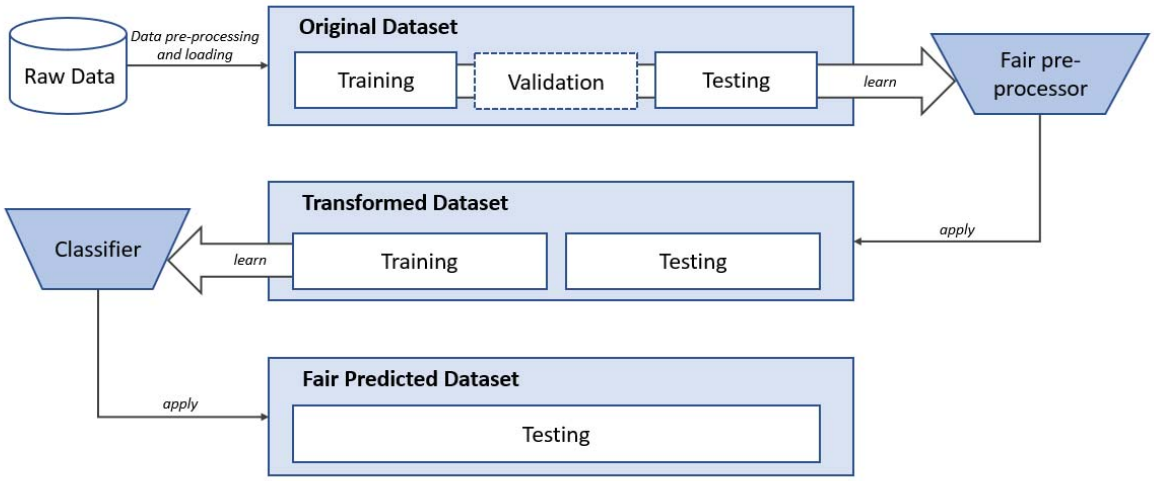
In order to algorithmically mitigate biases, three main approaches can be applied focusing on three different aspects of the model development chain. Pre-Processing, in-processing, and post-processing aim to achieve fairness and are more deeply explained in the following sections.

### 7.2.1. Pre-Processing

In-Processing focuses on modifying the training data so that a model can be trained with a fair data set to achieve fair results. Only if the algorithm is permitted to change the training data, then preprocessing can be applied. As shown in Figure 6 (adapted from Ballamy et al., 2018) a fair pre-processor is learned from the original data set from which a transformed dataset is created. Thereafter, a classifier learns which application results in a fair predicted dataset.

**Figure 6**

*Pre-Processing Process (adapted from Bellamy et al., 2018)*




Kamiran and Calders (2012) compared the efficiency of four bias mitigation algorithms in their paper naming suppression, massaging, reweighting, and sampling. Suppression is the sole deletion of the discriminating variable that was proven to be inefficient due to the redlining effect explained in section 3.3.2. Kamiran and Calders (2012) appraised the massaging approach whereby the labels of certain individuals are changed. The selection which labels to change is based on a ranker and is oriented towards the approach by Kamiran and Calders (2009). First, candidates closest to the decision border of belonging to the other class are relabeled resulting in only minimal accuracy downturns. That way discrimination is debilitated as some labels are modified; however, the overall class distribution remains unchanged. This approach has been appointed to be the most effective in achieving the highest fairness (Kamiran & Calders, 2012).

Preferential Sampling (PS) indicates comparable high efficiency likewise, an approach where borderline objects, meaning individuals being on the edge between two outcomes, are being assigned a high priority for duplication or deletion based on a ranker's estimation. Thereby, the size of a group is decreased by removing individuals closest to the decision line. In contrast, the sample size is increased by duplicating other data points which are also closest to the boundary based on a ranked listing. The procedure is repeated until the requested number of objects is achieved (Kamiran & Calders, 2012).

Lastly, reweighting does not involve changing labels but assigning weights to certain tuples. As the other two approaches interfere deeply into the original dataset, this approach is less intrusive. Instead of relabeling data, different weights will be assigned to them. Once again, the goal is discrimination reduction while perpetuating the same class probability. Thereby, lower weights will be assigned to objects that have been advantaged and disadvantaged treatment (Kamiran & Calders, 2012).

Another approach is provided by Zemel, Wu, Swersky, Pitassi, & Dwork (2013) who translate fairness into an optimization problem to achieve an efficient representation of the data with two opposing targets: to encode the data as effectively as possible while obscuring protected information.

Feldman et al. (2015) developed another bias mitigation technique focusing on mitigating disparate impact. Their results indicate a better fairness-utility trade-off than comparable techniques.

Calmon, Wei, Vinzamuri, Ramamurthy, and Varshney (2017) approach is closest to the one of Zemel et al. (2013) which is probabilistic framework for a flexible pre-processing approach allowing multiple measures and constraints. It allows, for instance, the possibility of multivariant, non-binary protected variables as well as the direct control of individual fairness. Thereby, an optimization problem is constructed that is trading off data utility, discrimination, and individual distortion.

### 7.2.2. In-Processing

In-processing focuses on finding a fair classifier to achieve algorithmic fairness without modifying the training data itself. Only if the algorithm is permitted to alter the learning process for a machine learning model, in-processing can be applied. As depicted in Figure 7 (adapted from Bellamy et al., 2018) a fair classifier is created from the original training data which leads to a fair predicted data set.

**Figure 7**

*In-Processing Process (adapted from Bellamy et al., 2018)*


Kamishima et al. (Kamishima, Akaho, & Sakuma, 2011; Kamishima, Akaho, Asoh, & Sakuma, 2012) named prejudice, underestimation, and negative legacy to be the three main sources for AI biases that are not mutually exclusive. Their focus was drawn to prejudice in which respect they developed a so-called “prejudice remover” which is a regularizer to expurgate prejudice. This regularizer is, therefore, independent of the determination of sensitive information and can be integrated into logistic regression models.

Besides, Zhang, Lemoine, and Mitchell (2018) followed the concept of bias mitigation developed by Hardt et al. (2016) and enhanced by Beutel et al. (2017). They developed four definitions and presented an in-processing approach in which respect they developed three definitions; demographic parity, equality of odds, and equality of opportunity. The demographic parity, also called statistical parity, concept is built on the perception that a decision must be made independent from protected attributes (Hardt et al., 2016). A predictor meets the stipulation of demographic parity if the probability for a positive outcome is the same for all values of the protected attribute. Equality of odds is achieved if a predictor and the protected attribute are conditionally independent

given the real true label. That is to say that that for all Zhang et al. (2018) technique aims to train accurate predictors while complying with these three equality definitions.

**Table 4**

*Overview Demographic Parity, Equality of Odds, Equality of Opportunity (adapted from Zhang, Lemoine, & Mitchell, 2018)*

| Definition | Explanation | Formula* |
|---|---|---|
| **Demographic Parity** | Definition is satisfied if the predictor and protected variable are independent. | $P(\hat{Y} = \hat{y})$ is equal for all Z: $P(\hat{Y} = \hat{y}) = P(\hat{Y} = \hat{y} \mid Z = z)$ |
| **Equality of Odds** | Definition is satisfied if the predictor and protected variable are conditionally independent given the output. | $P(\hat{Y} = \hat{y})$ is equal for all Z: $P(\hat{Y} = \hat{y} \mid Y = y) = P(\hat{Y} = \hat{y} \mid Z = z, Y = y)$ |
| **Equality of Odds** | Definition is satisfied with respect to a certain class y if the predictor and the protected output are independent conditioned on the output is equal to the class. | $P(\hat{Y} = \hat{y})$ is equal for all Z: $P(\hat{Y} = \hat{y} \mid Y = y) = P(\hat{Y} = \hat{y} \mid Z = z, Y = y)$ |

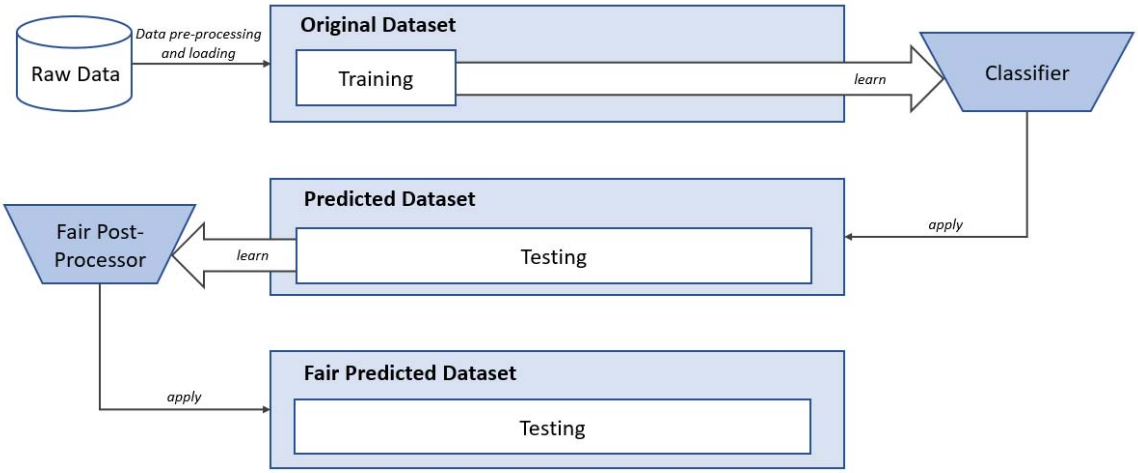
*Variables (input, output, predictor, protected variable) refer to ($X$, $Y$, $\hat{Y}$, $Z$)

### 7.2.3. Post-Processing

A post-processing approach is used if the learned model is a so-called black box model and cannot modify the training data or the learning algorithm (Bellamy et al., 2018). As shown in Figure 8 (adapted from Bellamy et al., 2018) from a predicted dataset a fair post-processor is learned. The application of this post-processor results in a fair predicted data set.

**Figure 8**

*Post-Processing Process (adapted from Bellamy et al., 2018)*




An equalized odds post-processing approach created by Hardt et al. (2016) aims to optimally adapt learned predictors in such a manner as to eradicate the discrimination in accordance with the pre-defined criteria. In addition, the framework enhances incentives by moving the charges of poor classification of the protected group to the decision-maker which in response improves classification accuracy. Thereby, Hardt et al. (2016) approach does not involve changing the training process as it increases complexity in their respect. As an alternative, they add a post-learning step where a non-discriminating predictor is developed. To achieve the optimal predictors satisfying the pre-defined conditions a loss minimization function is used. Further, to achieve their goal of better predictions while achieving equalized odds, characteristics that capture the target more directly, regardless of its correlation with the protected attribute, are needed. A predictor for balanced odds deduced from a score relies on the pointwise minimum ROC curve between different protected groups. The creation of predictors that are accurate in all groups is promoted by Hardt et al. (2016). For instance, appropriate data collection or by predicting characteristics that are existing in all groups.

Furthermore, Pleiss et al. (2017) offer a straightforward post-processing algorithm that aims to retain predictive information for randomly generating inputs to ensure parity and maintain calibration. This approach was proven to be unsuccessful as the post-processed predictions were fair, however, they were lacking its predictive power. In this respect, they generally concluded that calibration and error rates are contradicting targets that cannot be achieved simultaneously.

## 7.3. Learned Lessons and Limitations

The various measures presented in this section are the most prevalent ones in research and focus solely on quantifying fairness in supervised models by mostly assessing classification problems. Therefore, they solely provide guidance for regression problems and do not give any guidance for unsupervised learning (e.g. clustering or profiling problems). As ML becomes more prevalent in all areas, it is essential to have mature fairness assessment approaches for all types of algorithms. Fairness in an unsupervised setting is therefore not well-studied as it is more difficult to define. While in supervised learning, predictions can be mapped to decisions, in unsupervised learning no

predictions are made and therefore, no predictions can be related to any of the outcomes (Olfat & Aswani, 2019). Fairness measurements and mitigation approaches as from Dwork et al. (2012), Zemel et al. (2013), or Fedman et al. (2015) require the usage of labeled data which does not apply to unsupervised learning models. A further reason is that it is not apparent why fairness is an issue of concern for unsupervised learning, as no predictions are being made (Olfat & Aswani, 2019).

However, fairness in unsupervised learning is important and should not be neglected. First, unsupervised learning is applied to gain qualitative insights which can consist of dimensionality reduction to visualize high-dimensional data or to identify patterns through data clustering (Olfat & Aswani, 2019). The insights attained through these techniques can use data from protected groups and can then result in policies generated based on these insights (Olfat & Aswani, 2019). Second, ML methods often include unsupervised learning for the pre-processing phase as for instance when using dimensionality reduction for clustering. Most relevant papers and techniques developed to mitigate biases and achieve fairness for the data transformation step, require labeled data and cannot be used for unsupervised data transformations (Olfat & Aswani, 2019; see also Dwork et al. 2012; Zemel et al. 2013; Feldman et al. 2015).

Table 5 was developed to illustrate the unbalance of research conducted and papers published in the field of algorithmic fairness for unsupervised learning when compared to supervised learning. With almost 400 million publications and academic articles, Google Scholar is the world's largest database of its kind (Crew, 2019), and, as such, it was used to conduct this search exploration. Various search terms in the context of fairness of ML models in combination with supervised and unsupervised learning are being assessed. As illustrated in Table 5, a proportionally higher number of search results for "fairness" in the context of "supervised learning" than for unsupervised learning can be found. For most of the search terms, there are roughly double the number of publications concerning supervised learning than unsupervised learning. These findings are in line with the claim by Olfat and Aswani (2019) elaborated in the previous paragraph.

**Table 5**

*Comparison of the number of Search Results in Google Scholar (scholar.google.com) for Supervised Learning and Unsupervised learning search terms (status as of June 21, 2020).*

| Search Term for Supervised Learning | Number of Results | Search Term for Unsupervised Learning | Number of Results |
|---|---|---|---|
| "supervised learning" "fairness" | 6.780 | "unsupervised learning" "fairness" | 3.510 |
| "supervised learning" "fairness" "discrimination" | 1.850 | "unsupervised learning" "fairness" "discrimination" | 780 |
| "supervised learning" "fairness" "bias" | 3.510 | "unsupervised learning" "fairness" "bias" | 1.650 |
| "supervised learning" "fair" "bias" | 18.300 | "unsupervised learning" "fair" "bias" | 9.760 |
| "supervised machine learning" "fairness" | 1.390 | "unsupervised machine learning" "fairness" | 561 |
| "supervised machine learning" "bias" "fairness" | 764 | "unsupervised machine learning" "bias" "fairness" | 308 |

As prominent cases of algorithmic fairness concern classification cases like college admission and loan granting, biases in regression models should not be neglected. Previous research in this dissertation has shown that protected data can also be used in a regression context. For instance, the calculating of insurance payment fees, interest rates, or cancer mortality rate are examples where regression models can be used.

Next, a comparison in the search terms for classification and regression models is illustrated in Table 6. The difference in number of search results implies a gap between the research conducted in the field of fairness in classification versus regression models. For various search terms, more publications regarding the research in classification models can be found.

**Table 6**

*Comparison of the number of Search Results in Google Scholar (scholar.google.com) for classification and regression search terms (status as of June 21, 2020).*

| Search Term for Classification | Number of Results | Search Term for Regression | Number of Results |
|---|---|---|---|
| "classification" "fairness" | 387.000 | "regression" "fairness" | 212.000 |
| "classification" "fairness" "discrimination" | 86.700 | "regression" "fairness" "discrimination" | 42.100 |
| "classification" "fairness" "bias" | 114.000 | "regression" "fairness" "bias" | 99.400 |
| "classification" "fair" "bias" | 536.000 | "regression" "fair" "bias" | 439.000 |
| "classification" " algorithmic fairness" | 789 | "regression" " algorithmic fairness" | 472 |
| "classification " discriminating" | 392.000 | "regression " discriminating" | 274.000 |

Further, the fairness metrics and mitigation techniques evaluated utilize a narrow pool of datasets for their research. Various well-known scholars in this field like Kamiran & Kalders (2009, 2012), Feldman et al. (2015), Dwork et al. (2012), and Verma & Rubin (2018) base their prominent fairness definitions for classification on a German credit dataset (Lichman, 2013). It contains 20 attributes about 1000 loan applicants and the classification outcome: a good or bad credit score (Verma & Rubin, 2018). As this data set provides a good foundation to prove fairness measurements, it is still questionable to base findings solely on a single data set. Again, a classification example is applied and therefore it is questionable if these research findings can be translated to regression and unsupervised learning problems. In order to thoroughly analyze fairness assessment mitigation measures, it is crucial to use high bandwidth of data from different real-life examples.

# 8. SURVEY

As the specific sources and mitigation techniques of biases in AI systems are still widely disputed, we aim to address the general awareness of the issue. Awareness is the first step to all kinds of source identification and mitigation techniques, as individuals have to be aware of the risk to initiate the cautious creation, auditing, and application of fairer AI systems in the first place. Individuals from a multinational technology consultancy who develop and use AI systems are therefore asked to judge risk of AI systems being biased. As AI systems are embedded more and more in our everyday life, we also had people with no direct professional connection to AI systems to answer the survey. The main part of this survey includes a risk assessment of AI applications of various areas to see how individuals evaluate the risk of bias of AI decisions as opposed to decisions made by humans. Further, we aim to analyze if biased AI training and more diverse teams result in a higher risk assessment.

## 8.1. HYPOTHESIS DEVELOPMENT

The survey was conducted to test the following formulated hypothesis.

As the biases of AI systems are hard to assess for humans our first hypothesis is that that individuals perceive AI systems as fairer and less biased than humans. People have less knowledge about how AI systems process data in general and how biases can be incorporated into the systems. On the contrary, they know how biases can influence human decisions as they have experienced it in their own individual decision-making. Therefore, we expect that people generally underestimate the risk of the biased AI systems as they attribute a higher power to the system where decisions are conducted in a very analytical manner.

> *H1: Individuals perceive AI systems as fairer and less biased than humans.*

As lack of knowledge on bias related topics can result in a lack of awareness about the issue of Biased AI, with our second hypothesis, we aim to test if training increases the awareness of the risk of Biased AI systems. As the respondents have to classify the risk for biases of various scenarios of AI usage, we aim to analyze if respondents who received training on the issue classify the perceived risk of bias comparable higher than if respondents without any training do.

> *H2: Training on Biased AI / algorithmic fairness will result in higher awareness for the risk of Biased AI systems.*

Hypothesis 3 was formulated in order to test if the daily interaction with different minority groups within teams at work or at university will result in higher awareness. The survey respondents are, therefore, asked to specify the diversity of their teams at work.

> *H3: Diverse teams in the workforce will result in higher awareness of the risk of Biases in human and AI decision-making.*

## 8.2. Survey Methodology

The survey is composed of four main parts. The first part is the introduction to the survey, explaining the conditions, and asking for general personal information. The second part is the main part and contains multiple scenarios about human decision making and AI tools usage. The subsequent third part asks the respondent to assess multiple more general issues on biased AI. Last, the respondent is asked to answer a few last questions on team diversity, training on biased AI, and work experience in relevant fields.

### 8.2.1. Survey Introduction

The first part of the survey is a welcome message explaining the participation conditions and time required to take the survey (Appendix A). Throughout the survey, illustrations like the one in the appendix are used instead of simple text to catch the participants' attention and to represent the information in a more comprehensible form.

The following demographic information was asked to the respondents in the following order: age (Q1.1.), gender (Q1.2.), nationality (Q1.3), occupation (Q1.4.), and academic background (Q1.5.).

The general instructions to the survey are as follows.

> *In this project, I am interested in your opinion about decisions made by AI (Artificial Intelligence) agents. For that purpose, in the following pages, you will be presented with a set of scenarios. The scenarios involve a decision to be made by either a human or an AI agent. In particular, I am interested in your evaluation of the risk of these scenarios resulting in biased outcomes.*

In the following block of the survey an illustration provided a definition of bias. As there are various ways to perceive and understand bias, it is essential to instruct the participants of one uniform definition. Therefore, the illustration (Appendix B) defines bias as prejudice against certain minority groups. Illustrations display the four major types of bias; race, age, disability, gender, and sexual orientation. Further, it is explained that prejudice can lead to unfavorable treatment of one group in comparison to the majority group. In addition, the respondents are briefed that a bias can be intended, direct, unintended, or indirect.

### 8.2.2. Scenario Questions

The next part of the questionnaire contains 14 question blocks. For each block, a scenario is presented followed by four questions regarding that scenario. Seven different scenarios were created with each having two variations: with a human and an AI tool as decision-makers. The scenarios were created on the basis of common AI applications, such as CV screenings, or potential future applications like the usage of AI tools to decide who gets granted a ventilator during the COVID-19 crisis. Table 7 lists all the seven different scenarios (S) with the two different versions (V) each shown in the survey.

**Table 7**

*Scenario survey question overview*

| | Sector | Decision Maker | Scenario |
|---|---|---|---|
| **S1V1** | Human Resources Sector | Human | An HR person decides which applicants to invite for a job interview based on the applicants' CVs, the job description, and past hires' profiles. |
| **S1V2** | Human Resources Sector | AI | An HR recruitment AI tool decides which applicants to invite for a job interview based on the applicants' CVs, the job description, and past hires' profiles. |
| **S2V1** | Financial Sector (Bank) | Human | A bank clerk decides who to grant a bank loan by assessing the candidate's creditworthiness based on an interview, the person's characteristics, and past lending experience. |
| **S2V2** | Financial Sector (Bank) | AI | An AI tool decides who to grant a bank loan by assessing the candidate's creditworthiness based on the person's characteristics, and past lending experience. |
| **S3V1** | Insurance Sector | Human | An insurance clerk decides the amount of a person's insurance premium based on the person's characteristics, certain predefined criteria, and other similar clients' premiums. |
| **S3V2** | Insurance Sector | AI | An AI tool decides the amount of a person's insurance premium based on the person's characteristics, certain predefined criteria, and other similar clients' premiums. |
| **S4V1** | Criminal Sector I (Police) | Human | A policeman decides in which districts of Boston officers should patrol today to make random body checks. He decides based on his past experiences, day conditions, and crime statistics. |
| **S4V2** | Criminal Sector I (Police) | AI | An AI tool decides in which districts of Boston officers should patrol today to make random body checks. It decides based on past experiences data, day conditions, and crime statistics. |
| **S5V1** | Criminal Sector II (Prison) | Human | A judge decides on a criminal's prison length based on the person's characteristics, criminal records, past sentence length decisions, and legislation. |
| **S5V2** | Criminal Sector II (Prison) | AI | An AI tool decides on a criminal's prison length based on the person's characteristics, criminal records, past sentence length decisions, and legislation. |
| **S6V1** | Healthcare Sector | Human | A medical doctor makes a judgment about a person's skin cancer by conceptualizing the best treatment methods and calculating the mortality rate. The doctor takes into consideration past patient records and medical facts. |
| **S6V2** | Healthcare Sector | AI | A skin cancer image recognition software makes a judgment about a person's skin cancer by conceptualizing the best treatment methods and calculating the mortality rate. The software takes into consideration past patient records and medical facts. |

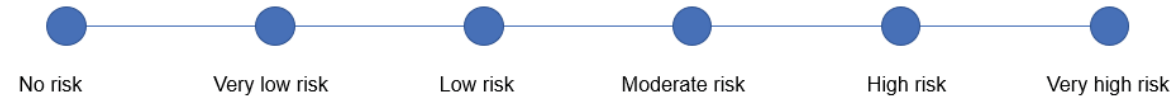
| | | | |
|---|---|---|---|
| **S7V1** | Healthcare Sector | Human | A health professional makes a judgment on which patient to grant a ventilator during the COVID-19 crisis. The health professional makes the decision based on the survival rates taking into consideration patients' medical records, experience with past COVID-19 cases, and medical facts. |
| **S7V2** | Healthcare Sector | AI | An AI tool makes a judgment on which patient to grant a ventilator during the COVID-19 crisis. The AI tool makes the decision based on the survival rates taking into consideration patients' medical records, experience with past COVID-19 cases, and medical facts. |

Each participant of the study is shown five different scenarios in one of their two versions which are randomly assigned by Qualtrics. The order of the four scenarios is also random. For each scenario four questions are asked. The first two questions asked are the same for all the scenarios. However, the following two questions have different possible answers: a version for the human scenario and a version for the decision based on an AI tool.

The first two questions focus on classifying the risk of a biased outcome and for that participants use a 6-points Likert scale (Figure 9) that goes from "No risk" to "Very high risk".

**Figure 9**

*Risk Assessment Likert Scale*

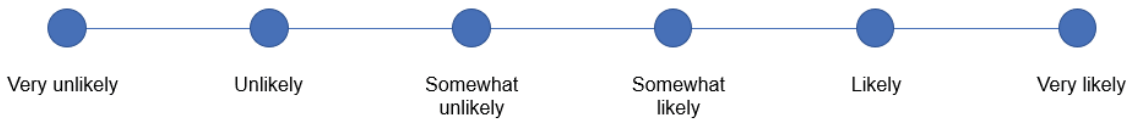



The first question is: "*What is the risk of this situation leading to a biased outcome?"* (Q2.1) and is the same for the human and AI scenario. The respondent is asked to use the first Likert scale (Figure 9) in order to classify.

The second question is: *"Next, indicate the risk of this situation leading to the following types of biases" (Q2.2.).* The respondent is asked to classify six different types of biases: race bias, gender bias, age bias, disability bias, sexual orientation bias. Further, he/she has the possibility to name another bias if desired. The respondent is asked to classify each bias with the same six options from the risk assessment Likert scale (Figure 9). These two questions are answered for all scenarios. The third and fourth questions depend on whether the version involves a human or AI system as the decision-maker. For the third and fourth questions of all scenarios, a likelihood Likert scale is used (Figure 10).

**Figure 10**

*Likelihood Likert Scale*




The third question focuses on the sources of potential bias: *"Indicate how likely is that the following sources are the cause of an observed bias." (Q2.3.)*

For a scenario involving the human decision-maker, seven possible sources are displayed to the respondent: Personal beliefs, experience level, education level, political view, socioeconomic background, and religious beliefs. Again, the respondent has the possibility to name another source if desired. Each of the sources is asked to be classified according to the likelihood Likert scale.
In contrast, when the respondent is given a scenario involving an AI tool as a decision-maker he/she is asked to classify the different sources according to the likelihood Likert scale (Figure 10). The first one is *"Programmers unintentionally/intentionally program their own biases and perspective into the AI systems."* focusing on the human bias of the people creating the algorithms. The second source provided is "*The AI systems are trained with biased, outdated, and inaccurate data."* indicating that past data incorporates human biases from past decisions. The third answer is "*The model is the reason as it uses the data in an inadequate way."* saying that the model has flaws leading to inadequate results. Further, the fourth answer states "*Inadequate applications of the AI systems."* determining that the false usage of the system results in biases.
Again, the respondent can add and evaluate another source if wanted. The purpose of this question is to understand the perception of the respondents for AI biases in a more detailed manner. The four named sources are chosen as they have been identified as the most common statements named in research and literature.

The fourth question is the following: *"For an observed bias, indicate how likely it is that the following measures would mitigate the bias." (Q2.4.).* Again, different options are provided depending on the decision-maker. For a human decision-maker, the following mitigation measures are listed and the participants is asked to use the likelihood Likert scale to rate each of the following options: "*Awareness training", "Quotas", "Filter sensitive information", "Having more diverse teams/ work environment", "Having an AI supportive system",* and "*Others, indicate which".*

For the scenarios with an AI tool as the decision-maker, the following proposed mitigation measures are listed: "*Stop using AI systems*", "*Cleaning data to remove bias",* "*Modifying the classifier*", "*Change the algorithm to promote fairer representation*", "*Change the usage and interpretation of AI models"* does and finally "*Others, indicate which" which* gives the respondents the opportunity to propose and evaluate other mitigation measures.

### 8.2.3. General Bias Questions

In this part, more general and not scenario-specific questions about biases are asked.

The first question proposed in this block aims to generally assess the respondents' perspective on the major source of bias. Thereby, a single answer question with four possible answers is presented as well as the option to say, "*I am not sure*". "*What is, in your opinion, the major source of potential biases against certain minorities in AI tools?"* (Q3.1.) has the following possible answers:

- Programmers unintentionally/intentionally program their own biases and perspective into the AI systems.
- The AI systems are trained with biased, outdated, and inaccurate data.
- The model is the reason as it uses the data the wrong way.
- AI systems are being used in the wrong way.
- I am not sure.

In the second question the respondent is required to decide between multiple bias mitigation approaches and choose the best one out of five bias mitigation techniques or indicate "*I am not sure"*. "*What is, in your opinion, the best approach to mitigate instances of biased AI*?" (Q3.2.) can be answered with:

- Stop using AI systems.
- Cleaning data to remove biases.
- Modifying the classifier.
- Change the algorithm to promote fairer representation.
- Change the usage and interpretation of AI models
- I am not sure.

The third question aims to analyze if the respondent is able to identify the redlining effect. The question is again a single answer question, however, with one specific correct answer. "*For the situations where the decisions were to be made by AI systems, what do you think will be the impact of not including protected features (e.g. gender, age, sexual orientation, disability, race) in the model or in the training data, in terms of biased outcomes?*" (Q3.3.) has the correct answer "*The bias will remain*.", the wrong answer "*The bias will be removed*" and "*I am not sure*". The first answer is correct as correlations can still cause biases (see section 3.3.2.).

### 8.2.4. Participants Additional Information Questions

The last part of the survey focuses on the respondent's workspace environment and working experience.

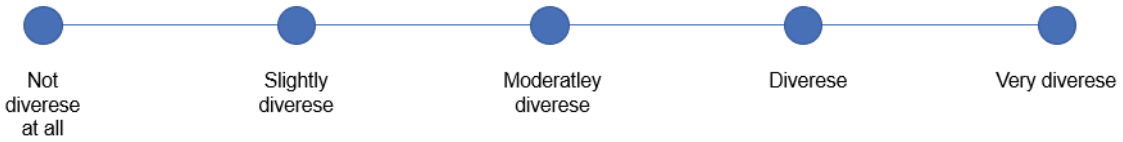
The first question is a single answer question asking if the respondent has completed any type of biased AI, algorithmic fairness, social bias, or data ethics training. This question provides insights if training on the issue results in higher risk awareness and expanded knowledge. "*Have you ever*

*completed training that covered topics on biased AI, algorithmic fairness, social bias, or data ethics?"* (Q4.1.) has the binary answers: yes and no.

Thereafter, the second question of this block is proposed aiming to analyze the diversity of the respondent's teams at work: "*How diverse is your team with respect to the following aspects?"* (Q4.2.). The diversity with respect to five minority groups is asked; age, gender, race, disability, and sexual orientation. The question is posed with a Likert scale (Figure 11) assessing the diversity on the basis of five options from not diverse at all to very diverse.

**Figure 11**

*Diversity Likert Scale*




The last question aims to classify the respondents' experience with data-driven projects and machine learning: *"How do you classify your experience with the following topics?" (Q4.3.)*

In this part, the respondent has to classify various experiences according to the experience level Likert scale displayed in Figure 12. The respondent has to classify his/her experience with "*Data-driven projects"* and *"Projects that involved the development of an AI/ML tool"*. Further, respondents have to classify if they have practical experience with AI/ Machine learning models developments such as "*Actively developing and programming AI/ML models",* and *"Actively preparing data for AI/ML models".* Besides, they have to assess their experience with AI fairness/ biases which was stated as *"Checking for algorithmic fairness/biases in AI/ML models"* in the survey. Further, we aim to know if the respondents have used individual-level data in projects before as this can result in critical biases. Therefore, the experience with *"Projects that involved AI/ML models used to make decisions about individuals using individual-level data"* has to be assessed.

The Likert scale used for all these experiences can be classified from "No experience" to "Expert experience" displayed in Figure 12.

**Figure 12**

*Experience Level Likert Scale*


## 8.3. EXECUTION

The survey was conducted from May 12th, 2020 to May 19th, 2020. Surveys that were 60% and above completed were used in the following analyses, leaving a total of 205 respondents. The

estimated time to complete the survey was seven to nine minutes. In fact, the median of the response time was 9,617 minutes.

The survey was shared through social networks like Facebook, Instagram, and LinkedIn; however, most responses were achieved through direct messages and private networks. Especially, through a network of prestigious universities in their field such as a German business school, a Portuguese business school, and Portuguese Information management school. Critically, part of the sample consisted of employees from an AI department of a multinational Fortune Global 500 service company.

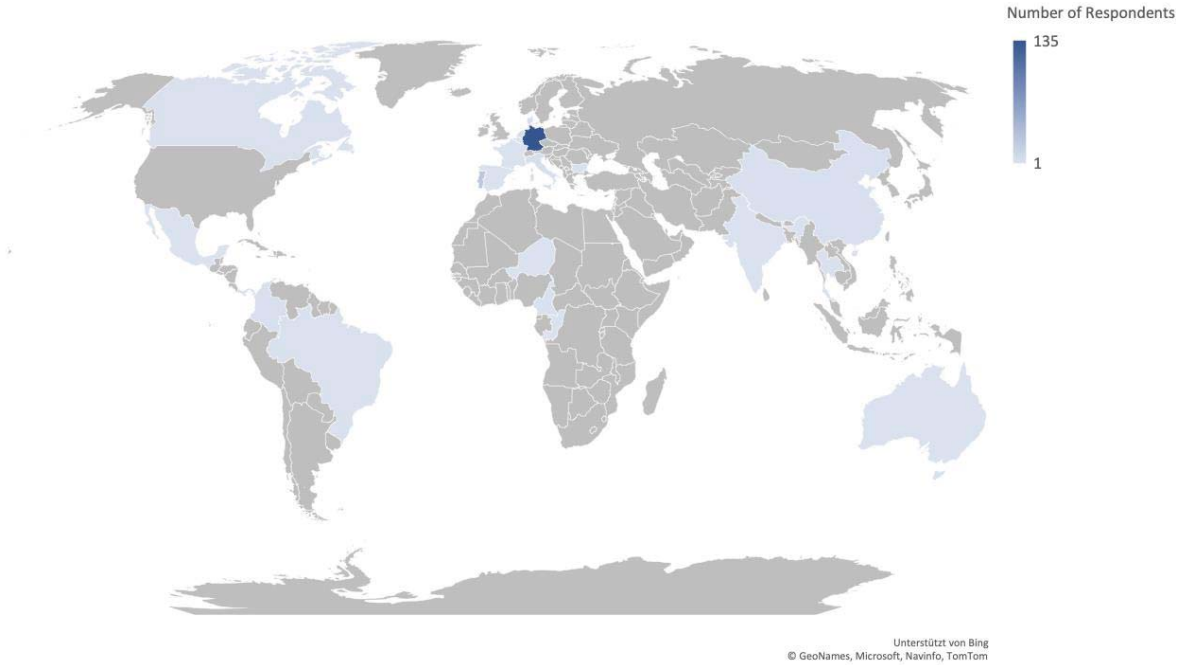
## 8.4. Descriptive Analysis

### 8.4.1. Profile Participants Analysis

From the respondents, 95 (46.34%) are female and 110 (53.66%) are male. Their ages vary between 19 and 58 with a mean age of 26.19 years. The respondents have 23 different nationalities as displayed in Figure 13. The highest proportion of respondents have German nationality with 135 German participants. The second highest stake of respondents is Portuguese which accounts for 27 participants.

**Figure 13**

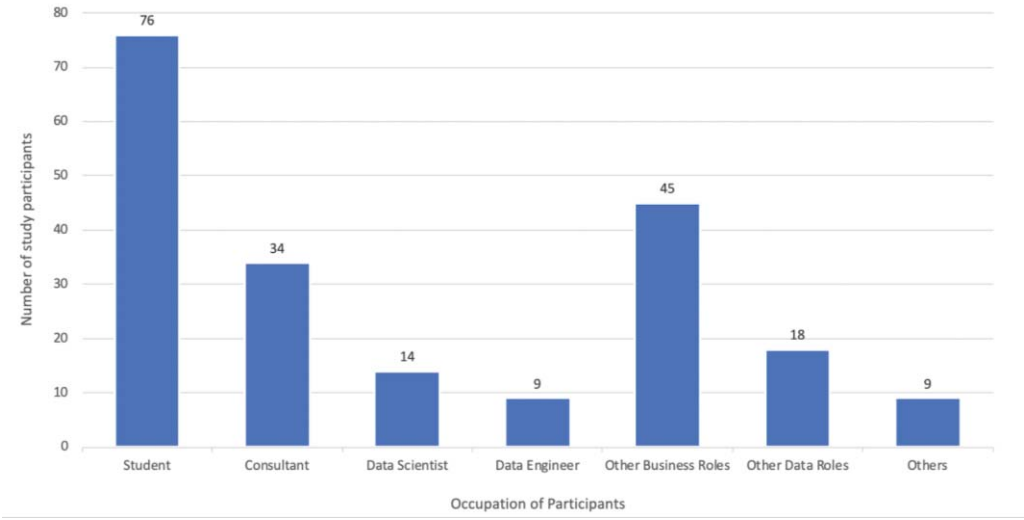
*Nationality Distribution of Respondents (Answer to Q1.3)*




To get a better understanding of the participants' knowledge and background, survey participants were asked to specify their occupation. From all respondents, 72 stated that "student" is their current occupation. Besides that, primarily people with labored in the business and technology field were questioned. In fact, 75 people have a business-related occupation such as Consultant, Human Resources, Marketing, Operations, or Finance. Technology-related occupations such as Data Scientist, Data Engineer, Product Owner, Machine Learning Engineer, or Software Engineer account for 38 participants.

**Figure 14**

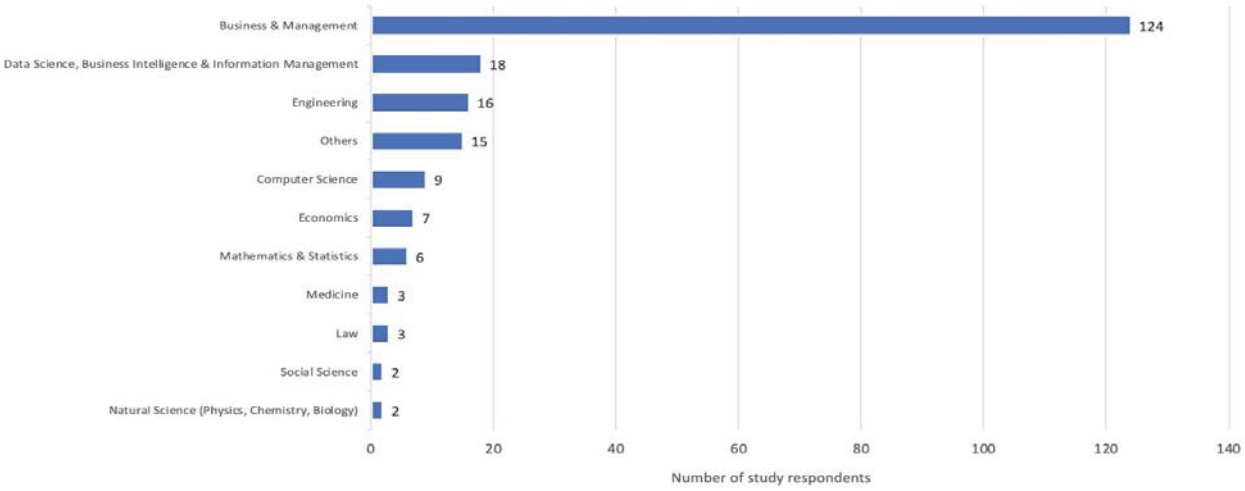
*Occupation of Survey Participants Grouped by Field of Work (Answer to Q1.4.)*




When asked about their educational background, 127 respondents have or are pursuing a degree in the business including similar denoted designations such as Management, Finance, and Accounting. In contrast, 36 respondents received a more technical education like Data Science, Information Systems, Engineering, or Mathematics. Other branches of education presented were Economics, Statistics, Communication, and other scientific fields including Physics, Medicine, Chemistry, and Biology.

**Figure 15**

*Academic Background of Participants Grouped by Academic Area (Answer to Q1.5.)*



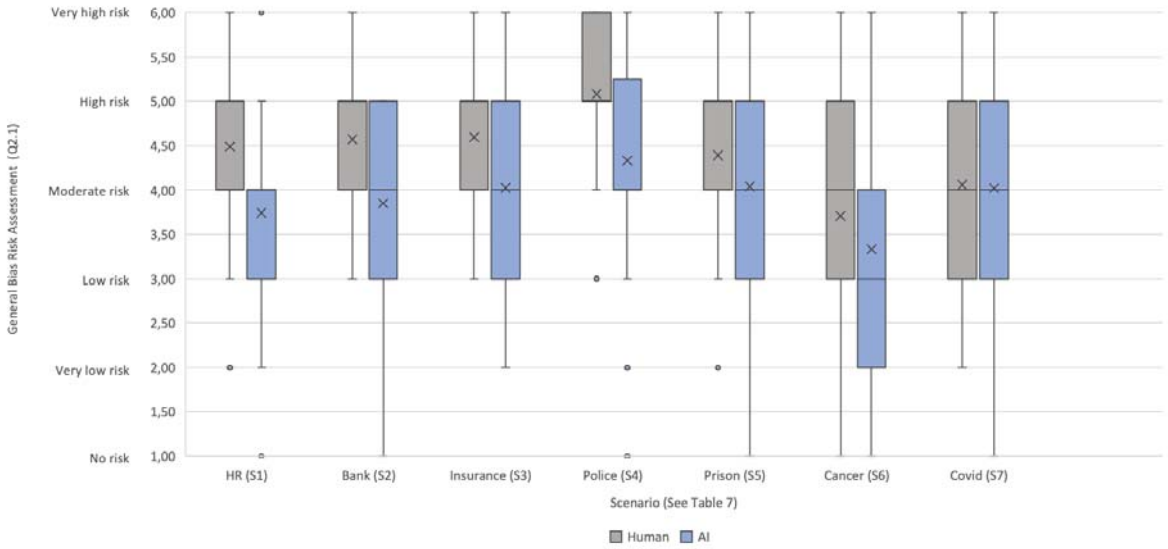

### 8.4.2. Risk Assessment of Scenarios

Figure 16 displays the difference in the risk perception of biases between AI models and human decisions. The average mean score for a human decision is 4.41 which translates into a moderate to high risk. In contrast, the mean risk of AI systems is 3.9 which corresponds to a moderate

risk. According to the survey, the scenario about a police officer patrolling and checking people (Table 7) poses the highest risk for both the AI systems (4.33) and the human decision-makers (5.08). The lowest risk was assigned to healthcare scenarios. The skin cancer recognition scenario had an average score of 3.33 for the AI system and 3.71 for the human scenario. To the COVID ventilator scenarios, a comparable equal risk score is assigned to the human and AI scenario (see Appendix C).

**Figure 16**

*Average Risk Assessment for each Scenario (Answer to Q2.1)*




In all the scenarios, the variance of the questions involving an AI system is on average substantially higher than the human scenarios. The variance for the human scenario lies between .77 and 1.15 while the variance for the AI scenario varies between 1.12 and 1.36 (see Appendix C). This implies that the assessment of the survey participants varied more for the AI scenarios than for the human scenarios where the answers provided by the different participants were more similar. As AI systems are black boxes and the data they are trained with as well as the creator are often unknown for the observer, the risk for biases is hard to measure. Therefore, a bigger discrepancy in the risk assessment of AI systems in comparison to human decision-making can be found as a result of the higher level of uncertainty in participants' responses.

**Table 8**

*Mean per scenario and bias*

| | | | Age Bias | Gender Bias | Race Bias | Disability Bias | Sexual Orientation Bias |
|---|---|---|---|---|---|---|---|
| **S1** | HU | Mean | 4.438 | 4.511 | 4.417 | 4.083 | 2.875 |
| | | Std | 0.848 | 0.975 | 1.028 | 1.318 | 1.383 |
| | AI | Mean | 3.848 | 3.391 | 3.043 | 3.174 | 2.587 |
| | | Std | 1.282 | 1.374 | 1.505 | 1.371 | 1.403 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **S2** | HU | Mean | 4.720 | 4.240 | 4.460 | 4.420 | 3.140 |
| | | Std | 0.882 | 1.153 | 1.403 | 1.197 | 1.440 |
| | AI | Mean | 4.083 | 3.458 | 3.771 | 3.542 | 2.362 |
| | | Std | 1.200 | 1.271 | 1.448 | 1.336 | 1.391 |
| **S3** | HU | Mean | 4.731 | 4.327 | 4.431 | 4.500 | 3.250 |
| | | Std | 0.952 | 0.944 | 1.209 | 1.365 | 1.329 |
| | AI | Mean | 4.512 | 4.143 | 3.837 | 4.619 | 2.953 |
| | | Std | 1.222 | 1.201 | 1.511 | 1.147 | 1.291 |
| **S4** | HU | Mean | 4.417 | 4.542 | 5.354 | 3.396 | 3.872 |
| | | Std | 0.919 | 0.988 | 0.758 | 1.300 | 1.146 |
| | AI | Mean | 3.950 | 4.390 | 4.854 | 2.951 | 3.000 |
| | | Std | 1.239 | 1.262 | 1.370 | 1.341 | 1.224 |
| **S5** | HU | Mean | 3.739 | 3.913 | 4.478 | 3.630 | 3.304 |
| | | Std | 1.163 | 1.347 | 1.329 | 1.404 | 1.482 |
| | AI | Mean | 3.500 | 3.540 | 4.020 | 3.120 | 2.740 |
| | | Std | 1.093 | 1.358 | 1.548 | 1.223 | 1.266 |
| **S6** | HU | Mean | 4.279 | 3.545 | 3.422 | 3.733 | 2.591 |
| | | Std | 1.182 | 1.337 | 1.390 | 1.452 | 1.227 |
| | AI | Mean | 3.583 | 2.917 | 3.229 | 2.979 | 1.957 |
| | | Std | 1.456 | 1.350 | 1.519 | 1.451 | 1.158 |

The average means for risk assessment for the race bias for each scenario is shown in Appendix D. In all the shown scenarios, the human mean risk is lower than the risk for biases than the AI tool. The significance of the difference is discussed in chapter 8.5.1. Besides, the risk assessment for the police scenario (S4V2: *m* = 4.33, *sd* = 1.36) and prison scenario (S5V2: *m* = 4.04, *sd* = 1.19), both examples from the criminal sector, pose the highest risks for a biased outcome in terms of the mean risk assessment. This is in line with the reality (see chapter 5.3.) as in the criminal sector, specific races are often being discriminated against due to various human biases. For instance, dark-skinned criminals receive more severe penalties than white perpetrators (Steffensmeier et al., 1998; Petersilia, 1983; Spohn, 1990).

The lowest risk is assigned to the cancer scenario (S6V1: *m* = 3.71, *sd* = 1.15, S6V2: *m* = 3.33, *sd* = 1.15, see Appendix C). The research from section 5.4. indicates, however, that risk for bias in this sector is substantial. For instance, prediction in mortality rates in cancer research performed by AI models vary in its accuracy as about 20% between the majority and protected group members (Chen, Johansson, & Sontag, 2018). Since these models are trained with mostly white male data (Boulamwini & Gebru, 2018), the model can predict skin cancer with better accuracy for the majority group. The results indicate that participants in our sample were not aware of this kind of risk in health-related scenarios.

Both race and gender biases, the biggest discrepancy between the human and the AI system can be seen in the scenario illustrating CV screening of applicants. In appendix D and E can be seen that an HR person inviting candidates to an interview poses a risk due to human biases that AI models which

decide who to invite to an interview based on the candidates' characteristics, past hiring experience, and job description pose a risk as well.

In general, participants assessed the discrimination against certain genders high compared to other protected variables. As the sample is balanced in terms of gender (46.34% are women), it can be concluded that the participants relate to gender bias comparable more. In addition, the working teams of the participants are the most diverse in terms of gender, with an average classification as moderately diverse, making this type of bias more relatable and apparent.

For the protected variable age, the risk is on average moderate. In contrast to race, the scenarios from the criminal sector (S4, S5) are assigned comparable lower risk. The COVID scenario (S7) was classified as having a high risk for age bias which can be explained by the fact that ventilators are in fact assigned to younger patients as they have higher survival rates (Fox, 2020).

In Appendix G the mean of the risk for disability bias for each scenario is displayed. The risk is classified as comparable low as opposed to the other protected variables. Especially in the criminal sector, the risk for disability bias is rated as a low risk which was graded as high risk for the other protected variables. However, for the insurance and COVID scenarios, the risk rates are comparatively high, which is reasonable since people with disabilities have lower survival rates (Dickinson & Kavanagh, 2020; Unicef, 2020).

Sexual orientation as illustrated in Appendix H (22) achieved the overall lowest rating with respect to risk for bias with an average low-risk grading. In the police patrolling scenario, the risk of human bias is graded as comparable high.

### 8.4.3. Influence of Individual Aspects on Scenario Outcomes

Figure 17 shows a comparison of the risk assessment between men and women. Thereby, no significant difference can be found as the p-value is not lower than significance level (Table 9).

**Figure 17**

*Average Risk Assessment for Men and Women (Answers to Q2.1. grouped by Q1.2.)*

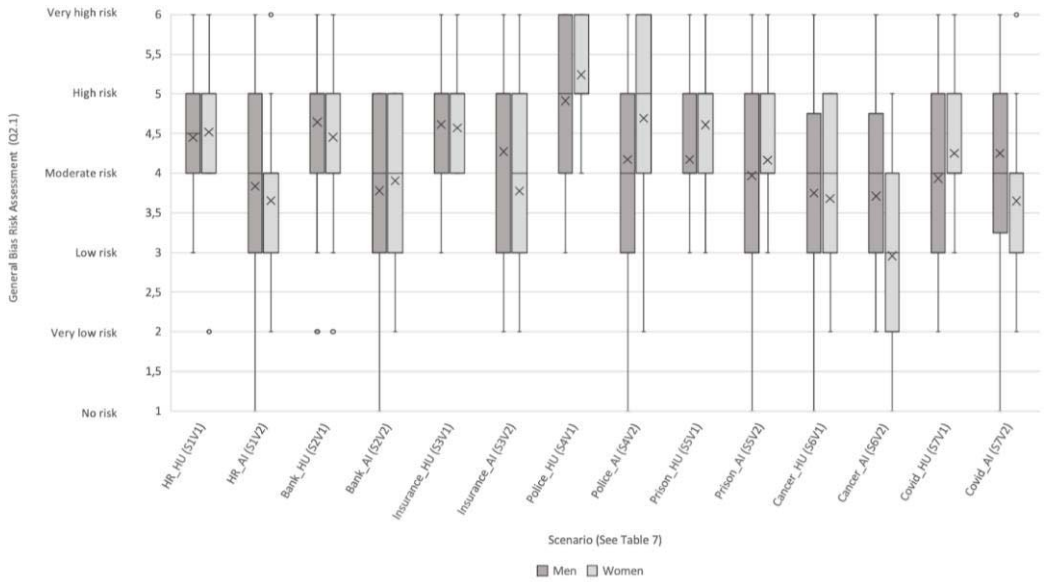


**Table 9**

*Independent Samples T-Test (Male unequal female)*

| | t | df | p | Mean Difference | SE Difference |
|---|---|---|---|---|---|
| **HR_HU** | -0.289 | 47 | 0.774 | -0.067 | 0.233 |
| **HR_AI** | 0.511 | 45 | 0.612 | 0.181 | 0.355 |
| **Bank_HU** | 0.624 | 48 | 0.536 | 0.188 | 0.302 |
| **Bank_AI** | -0.387 | 46 | 0.700 | -0.127 | 0.328 |
| **Insurance_HU** | 0.167 | 50 | 0.868 | 0.041 | 0.248 |
| **Insurance_AI** | 1.321 | 42 | 0.194 | 0.500 | 0.379 |
| **Police_HU** | -1.494 | 46 | 0.142 | -0.327 | 0.219 |
| **Police_AI** | -1.154 | 40 | 0.255 | -0.520 | 0.451 |
| **Prison_HU** | -1.430 | 44 | 0.160 | -0.435 | 0.304 |
| **Prison_AI** | -0.558 | 48 | 0.579 | -0.198 | 0.354 |
| **Cancer_HU** | 0.210 | 46 | 0.834 | 0.071 | 0.340 |
| **Cancer_AI** | 2.356 | 46 | 0.023 | 0.750 | 0.318 |
| **Covid_HU** | -1.151 | 48 | 0.256 | -0.317 | 0.275 |
| **Covid_AI** | 1.640 | 46 | 0.108 | 0.600 | 0.366 |

*Note.* Student's t-test.

Table 10 comprises the average risk assessment, mode and standard deviation group by educational backgrounds. An independent sample t-test has shown that no significant difference between an educational background from STEM (science, technology, engineering and mathematics) and business background can be found besides the police scenario including a human decision maker (Appendix I).

**Table 10**

*Average risk assessment per educational background (Answers to Q2.1. grouped by Q1.5.)*

| Educational Background | N (Number of respondents) | Scenario | Mean | Standard Deviation |
|---|---|---|---|---|
| **Business** | 131 | Human | 4.40517241 | 1.04428727 |
| | | AI | 3.86098655 | 1.28870803 |

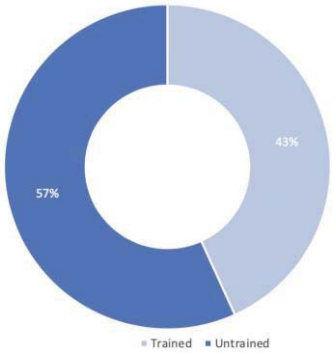
| | | | | |
|---|---|---|---|---|
| **Computer Science** | 8 | Human | 4.44444444 | 1.01379376 |
| | | AI | 3.8 | 1.13529242 |
| **Data Science** | 8 | Human | 4.07692308 | 1.03774904 |
| | | AI | 4.28571429 | .72627304 |
| **Engineering** | 15 | Human | 4.38095238 | .66904338 |
| | | AI | 4.47058824 | .94324222 |
| **Information Systems** | 10 | Human | 4.35 | .87509398 |
| | | AI | 3.47368421 | 1.34859684 |
| **Mathematics** | 6 | Human | 4.92857143 | 0.73004591 |
| | | AI | 3.81818182 | 1.32801972 |
| **Others** | 27 | Human | 4.44117647 | 1.25989842 |
| | | AI | 3.93939394 | 1.14398956 |

### 8.4.4. Biased AI Training

In the survey participants were asked if they have ever completed training that covered topics on biased AI, algorithmic fairness, social bias, or data ethics. Figure 18 shows that 43% of the survey respondents have participated in this kind of training before.

**Figure 18**

*Trained and Untrained Participants in Terms of Biased AI Training (Answers to Q4.1.)*

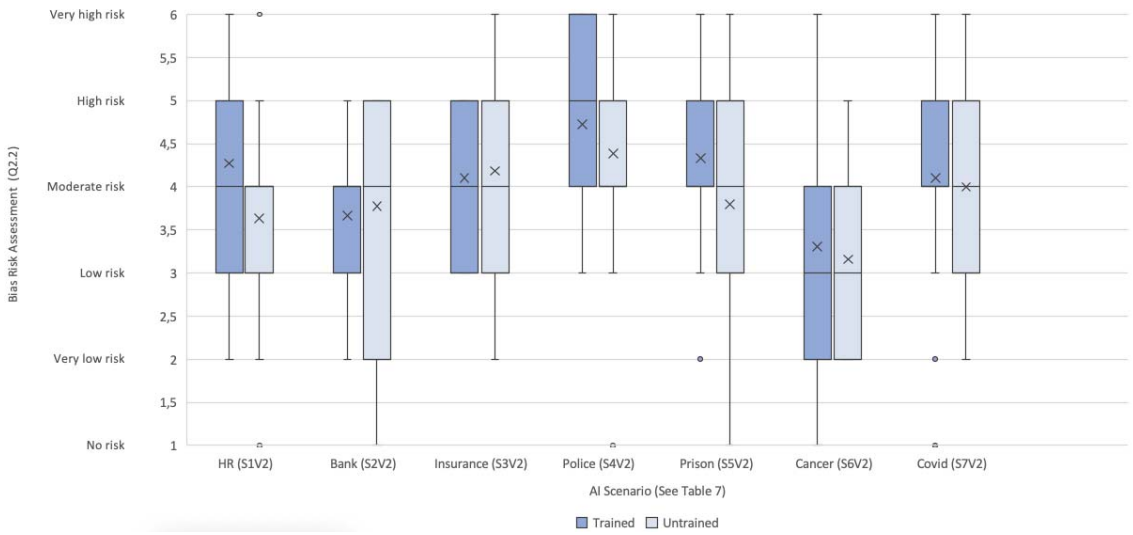



*Note. Trained participants answered "yes" and untrained answered "no"*

Figure 19 displays the risk assessment of the scenarios for a trained person meaning a person who responded "yes" to the question if he/she ever completed training on that issue whereby an untrained participant is a person who did answer "no".

**Figure 19**

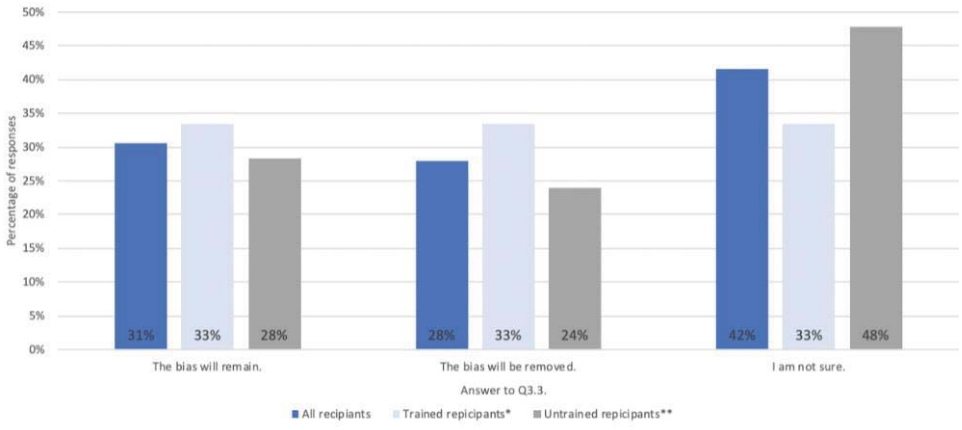
*Box Plot of Risk Assessment for Trained and Untrained Participants (Answers to Q2.1. grouped by Q4.1.)*




The next graphic shows the answers to the question "What do you think will be the impact of not including protected features in the model or in the training data, in terms of biased outcomes?". Here, the right answer is "The bias will remain." as correlation can still cause biases (see section 3.3.1.). Trained participants include the people who answered the question "Have you ever completed training that covered topics on biased AI, algorithmic fairness, social bias, or data ethics?" with a "Yes". However, even if respondents were trained on the topic biased AI, they provided a wrong answer more often than untrained people. This can be explained by the fact that untrained participants provided the answer "I am not sure" 48% of the time in comparison to just 33% of the untrained people. This implies that trained participants feel more secure about their knowledge and therefore rather provide wrong answers than acknowledge that they are not sure about the answer. Further, this observation indicates that the training participants received was not sufficient as only 5% more correct answers were achieved by trained participants.

**Figure 20**

*Redlining Effect Knowledge Question (Answer to Q3.3.) for all, Trained and Untrained Participants (Grouped by Q4.1.)*



### 8.4.5. Sources of Bias Analysis

Survey participants were asked if they could provide their assessment of the likelihood of various sources leading to a biased outcome.

For the scenarios involving a human decision-maker the personal beliefs, experience level, education level, political level, and socioeconomic background were assessed as possible sources of human bias. As Figure 21 displays, religious beliefs have been classified as the lowest likelihood to account for biased outcomes. For the scenario from the HR sector, personal beliefs were assigned the highest likelihood, followed by experience level, education level, and socioeconomic background. For scenarios from the banking and insurance sector, comparable assessments were made whereby socioeconomic backgrounds were assigned the highest likelihood followed by personal beliefs. The scenarios of the policeman patrolling the streets and the scenario involving a judge deciding on a prison sentence are performing similarly in terms of respondents' evaluation. Thereby, personal beliefs are classified as having the highest possibility of resulting in a biased outcome. Further, experience level, socioeconomic background, and political views were also classified between somewhat likely and likely resulting in a biased outcome. Education level and religious beliefs were both classified as less likely than the other sources. The last two scenarios, both examples from the healthcare system, perform similarly in their assessment. Thereby, the experience level of the health professionals was indicated as having the highest likelihood of resulting in biases. Political views and religious beliefs were assigned the lowest likelihood of all assessments conducted.

**Figure 21**

*Assessment of Sources of Biased AI for Scenarios with Human Decision-Maker (Answers to Q2.3)*

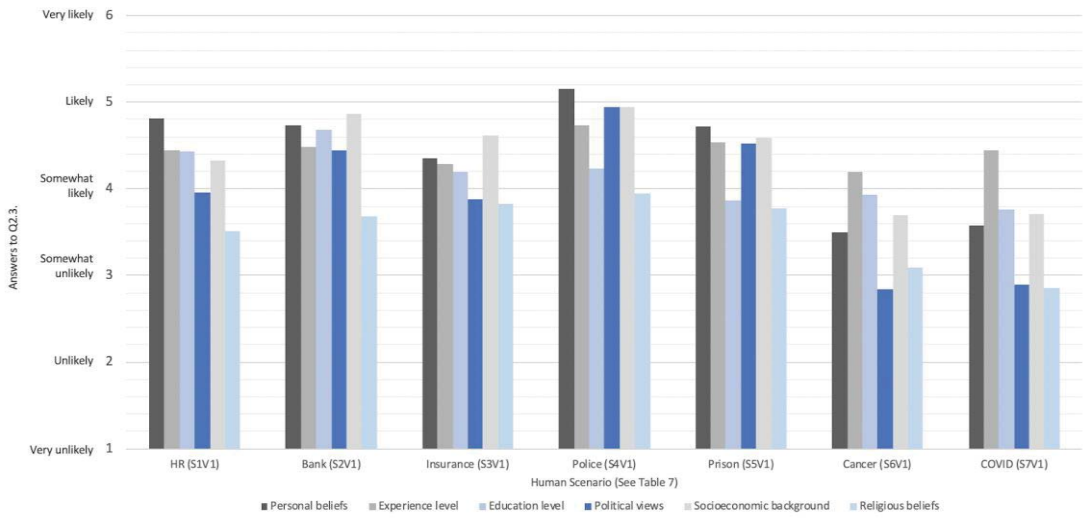
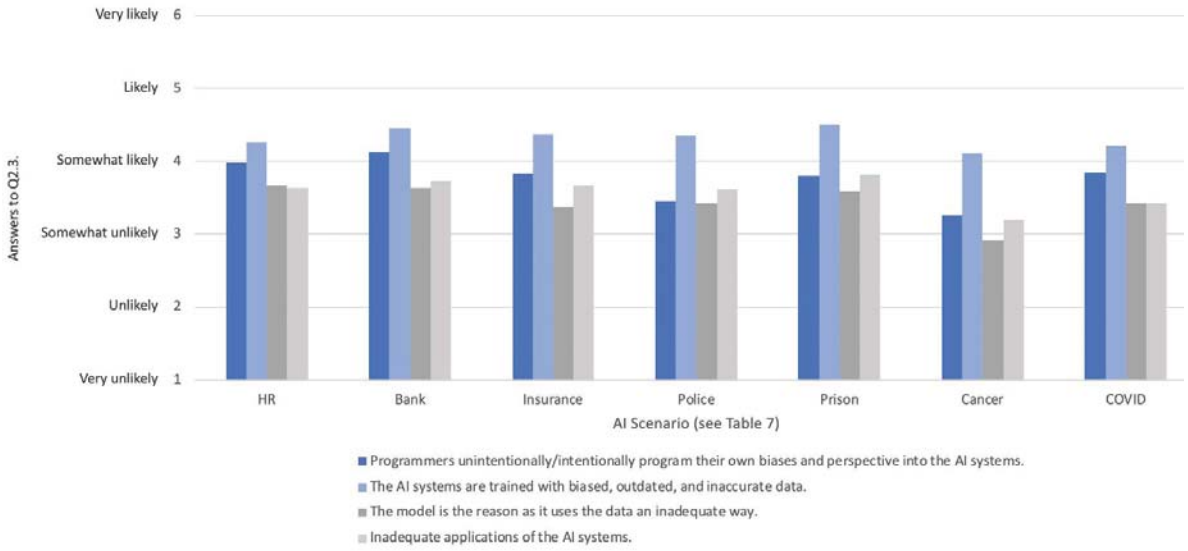



Figure 22 displays the same evaluation for the scenarios involving an AI system as a decision-maker. Four possible sources are proposed as leading to a biased outcome. The source “The AI systems are trained with biased, outdated and inaccurate data” was selected as the most plausible in all the scenarios.

**Figure 22**

*Assessment of Sources of Biased AI for Scenarios with an AI System as Decision-Maker (Answers to Q2.3)*




### 8.4.6. Bias Mitigation Techniques Analysis

Figure 23 displays the perception of bias mitigation techniques assessed by survey participants. From the set of mitigation techniques for the human biases' quotas are perceived as being generally much less efficient than the rest.

**Figure 23**

*Assessment of Bias Mitigation Techniques for Human Decision-Makers Scenarios (Answers to Q2.4)*

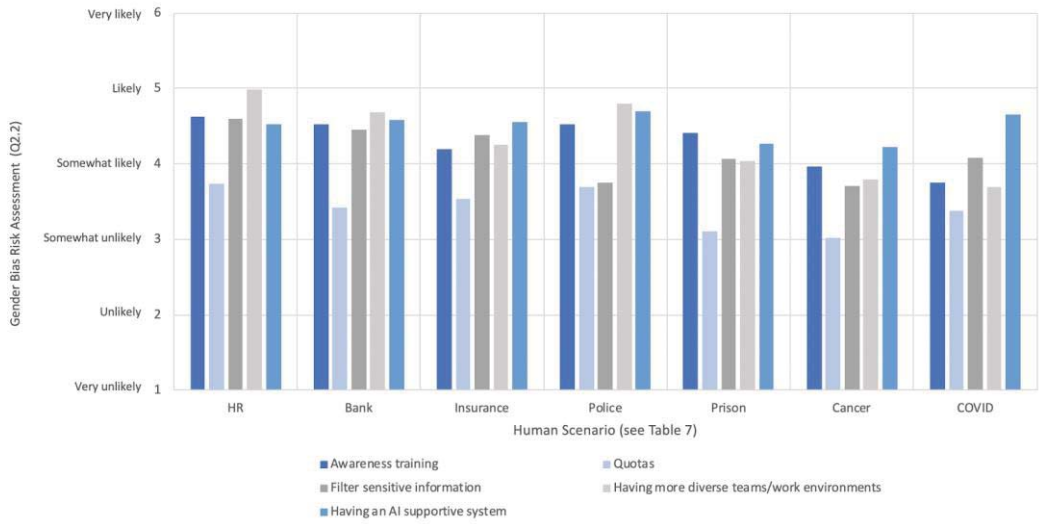
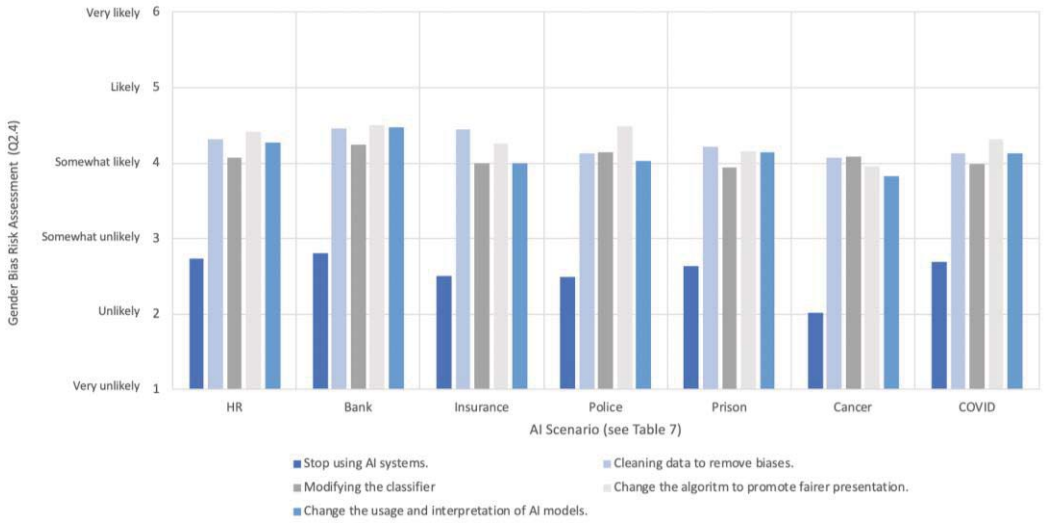



Figure 24 illustrates mitigation techniques for AI biases. As one can see "stop using AI systems" is seen as the least likely mitigation technique to be able to remove biases. The four other mitigation techniques proposed are achieving similar promising results as they are classified as on average "likely" to remove biases.

**Figure 24**

*Assessment of Bias Mitigation Techniques for Scenarios with AI Systems as Decision-Makers (Answers to Q2.4)*

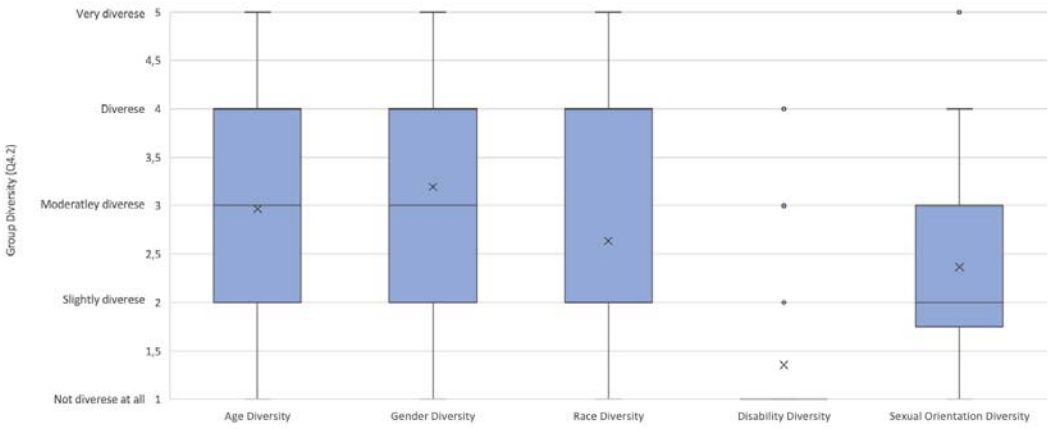



## 8.4.7. Team Diversity Analysis

Respondents were asked to classify the diversity of their team at work or at university. As this question belongs to the final part of the survey, only 118 of the 205 respondents were recorded from this part. The tasks included grading the team diversity in respect to race, gender, age, disability, and sexual orientation according to a diversity Likert scale (Figure 11). The chart below (Figure 25) displays the classification provided by the respondents. The discrepancy is particularly apparent for the feature “disability” where 91 respondents stated that their team is not diverse at all. In terms of average diversity, sexual orientation comes second whereby 74 people classified this kind of diversity as not diverse at all or just slightly diverse. Third least diverse comes the race in terms of their average diversity score. Most people graded their team diversity as just slightly diverse. 77 people are specified to have a at least moderate diverse team in terms of age. Gender diversity performs the best in this survey as most people classified their team as being diverse.

**Figure 25**

*Diversity of Teams of Respondents (Answers to Q4.2.)*



### 8.4.8. Influence of AI/ML Experience on AI Risk Assessment

In order to analyze the relation between a person's experience with AI systems and their submitted risk assessment, Pearson's correlations are analyzed in Table 11. The correlation between the risk assessment of certain scenarios is substantially low and does not exceed a correlation of .36. This correlation is also the only correlation reaching the significance level (p-value < .05). Even direct experience with algorithmic fairness checks and projects that involved individual-level data does not correlate substantially with the risk perception of biased AI models. It is hazardous that individuals actively engaging and creating AI models are not more risk averse than the average (even though this conclusion is an imprecise extrapolation, since it is based on a null result). An additional correlation between training and the experiences is analyzed in Table 11 and shows that individuals who actively prepare data for AI/ML models and who work on projects that involve AI/ML models using individual-level data are more likely to have participated in training on biased AI or similar topics.

**Table 11**

*Pearson's Correlation Between Scenarios (Q2.1.) and Experience (Q4.4.), Biased AI Training (Q4.1.)*

| *Variable* | | *Data-driven projects* | *Projects that involved the development of an AI/ML tool* | *Actively developing and programming AI/ML models* | *Actively preparing data for AI/ML models* | *Checking for algorithmic fairness/biases in AI/ML models* | *Projects […] using individual-level data* |
|---|---|---|---|---|---|---|---|
| **Biased AI (Q4.1.)** | Pearson's r | .253 | .238 | .290 | **.411** | .206 | **.348** |
| | p-value | .111 | .133 | .066 | ***.008** | .195 | ***.026** |
| **HR AI (S1V2)** | Pearson's r | .070 | .243 | **.360** | .274 | .119 | .052 |
| | p-value | .669 | .130 | ***.022** | .087 | .463 | .750 |
| **Bank AI (S2V2)** | Pearson's r | -.079 | -.164 | -.174 | -.052 | -.137 | -.177 |
| | p-value | .642 | .331 | .304 | .760 | .417 | .294 |
| **Insurance AI (S3V2)** | Pearson's r | -.200 | -.043 | -.030 | -.194 | -.161 | -.154 |
| | p-value | .243 | .804 | .864 | .257 | .349 | .369 |
| **Police AI (S4V2)** | Pearson's r | .084 | .205 | .284 | .225 | .243 | .121 |
| | p-value | .592 | .187 | .065 | .146 | .116 | .439 |
| **Prison AI (S5V2)** | Pearson's r | .035 | .229 | .012 | .086 | .113 | .040 |
| | p-value | .826 | .145 | .942 | .589 | .476 | .803 |
| **Cancer AI (S6V2)** | Pearson's r | -.090 | -.033 | -.041 | -.023 | .041 | -.045 |
| | p-value | .567 | .835 | .792 | .881 | .794 | .773 |

## 8.5. Statistical Hypothesis Tests

The hypothesis developed requires statistical testing to provide a quantitative foundation.

### 8.5.1. Statistical Test of Hypothesis 1

*H1: Individuals perceive AI systems as fairer and less biased than humans.*

In order to test H1, the risk perceptions of the AI and human scenarios are compared. Whitin inferential statistics, T-tests are usually used to compare means. There are paired t-tests, which are used to compare the mean of paired observations (e.g., same respondents answering both questions A and B), and independent t-tests, which assess difference between independent samples (different groups of respondents answering to questions A and B).
This survey, however, contains paired and independent samples as the assignment of five of the 14 scenarios (seven scenarios in two versions, AI or humans) was done randomly. This means that for some scenarios the same participants could have seen both the AI and the human versions, whereas for other scenarios only one of the two versions were presented to each participant. To apply an independent t-test, the paired observations would not be considered and vice versa if a paired test was to be used. To not exclude any of our observations, a different approach must be applied. Derrick, Toher, and White (2017) suggest an approach to solve exactly that matter. Derrick et al. (2017) suggest using an interpolation between the paired sample t-test and Welch's test. A more detailed descriptions of the mathematical equations of the overlapping sample t-test can be found in Appendix J. The calculations were, however, not conducted manually since the R package "Partiallyoverlapping" from Derrick, Toher, et al. (2017) was used. An example of the R code can be found in Appendix K. The results of the partially overlapping t-test are shown in Table 12.

**Table 12**

*Paired T-Test (in R), Means and Standard Deviation of General Risk Assessment*

| Scenario | | Mean | Sd | Statistic | Parameter | p-value | Estimate |
|---|---|---|---|---|---|---|---|
| HR | HU Unpaired | 4.406 | .837 | -3.934 | 56.736 | ***.00023** | -.745 |
| | HU Paired | 5.6 | .699 | | | | |
| | AI Unpaired | 3.8 | 1.349 | | | | |
| | AI Paired | 3.647 | .931 | | | | |
| Bank | HU Unpaired | 4.628 | 1.114 | -3.535 | 70.233 | ***.00073** | -.727 |
| | HU Paired | 4.4 | .910 | | | | |
| | AI Unpaired | 3.909 | 1.182 | | | | |
| | AI Paired | 3.667 | .976 | | | | |
| Insurance | HU Unpaired | 4.474 | .922 | -2.563 | 56.324 | ***.0131** | -.573 |
| | HU Paired | 4.929 | .616 | | | | |
| | AI Unpaired | 4.167 | 1.234 | | | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | AI Paired | 3.714 | 1.326 | | | | |
| Police | HU Unpaired | 4.947 | .733 | -3.251 | 50.92 | ***.00204** | -.75 |
| | HU Paired | 5.6 | .699 | | | | |
| | AI Unpaired | 4.218 | 1.431 | | | | |
| | AI Paired | 4.7 | 1.059 | | | | |
| Prison | HU Unpaired | 4.379 | 1.049 | -1.711 | 66.217 | .0918 | -.351 |
| | HU Paired | 4.412 | 1.064 | | | | |
| | AI Unpaired | 4.030 | 1.237 | | | | |
| | AI Paired | 4.059 | 1.144 | | | | |
| Cancer | HU Unpaired | 3.528 | 1.230 | -1.694 | 73.248 | .095 | -.375 |
| | HU Paired | 4.25 | .622 | | | | |
| | AI Unpaired | 3.361 | 1.199 | | | | |
| | AI Paired | 3.25 | 1.055 | | | | |
| Covid | HU Unpaired | 4.091 | .947 | -.289 | 62.519 | .774 | -.06 |
| | HU Paired | 4 | 1 | | | | |
| | AI Unpaired | 3.903 | 1.165 | | | | |
| | AI Paired | 4.176 | 1.468 | | | | |

*Note: *Significance level 0.05 is reached*

The statistics and corresponding p-values suggest significant differences in the risk assessment for the following scenarios: HR, bank, insurance, and police scenario and shows that the risk assessment of a human decision maker is substantially higher than the risk perceived when an AI system comes to a decision (see Table 12 for the means and standard deviations). In sum, for most of scenarios, our hypothesis was supported.

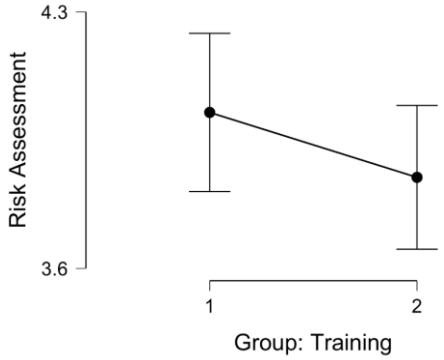
### 8.5.2. Statistical Test of Hypothesis 2

*H2: Training on Biased AI / algorithmic fairness will result in higher awareness for the risk of Biased AI systems.*

To test this hypothesis, an independent sample student's t-test is conducted to analyze the fairness assessments provided by trained and untrained participants. Even though the risk assessments are higher for the trained groups (group 1 in Figure 26, mean = 4.028) than for the untrained group (group 2 in Figure 26, mean = 3.848), suggesting they might be more aware of the biases in AI, our hypothesis was not supported, $t(280)$=1.185, $p$=.118, since the p value did not reach statistical significance.

**Figure 26**

*Descriptive Plot: Risk Assessment of Trained (1) and Untrained Study Participants (2)*




*Note: Shows mean of group 1 (4.026) and group 2 (3.848) and confidence interval (0.95)*

### 8.5.3. Statistical Test of Hypothesis 3

> *H3: Diverse teams in the workforce will result in higher awareness of the risk of Biases in human and AI decision-making.*

Table 13 displays the person's correlation between the survey participant's team diversity score of age, gender, race, disability and sexual orientation and the average risk score for all human and all AI scenarios. The p-value suggests that solely the correlation of race and disability diversity for the AI scenarios is significant. Thereby, it demonstrates that a higher race diversity as well as disability diversity within a group results in a more risk averse perception of the risk of AI bias.

The computed figures are reasonable as age diversity and gender diversity are more established in the workforce as has shown our survey (age diversity mean = 2.97, gender diversity mean = 3.19) and therefore can be the norm and are the baseline rather than a risk aversion trigger. Race diversity in teams has the highest significant impact on the average risk score for AI scenarios and correlates positively with risk aversion. Disability diversity has the lowest average risk score (disability diversity mean = 1.36) and shows a significant impact on the average risk score for AI scenarios (p-value = 0.034). Sexual orientation diversity has the second lowest diversity score (sexual orientation diversity mean = 2.36) whereby it has to be considered that sexual orientation of co-workers cannot be always evident to the survey respondents as sexual orientation does not have to be discernible. A more detailed table of Pearson's correlations between the risk assessment for each scenario and team diversity can be found in Appendix L.

**Table 13**

*Pearson Correlation Between the Average Risk Assessment and Team Diversity*

| | | Average Risk Score for Human Scenarios | Average Risk Score for AI Scenarios |
|---|---|---|---|
| Age Diversity | Pearson's r | -.056 | .070 |
| | p-value | .550 | .453 |

| | | | |
|---|---|---|---|
| Gender Diversity | Pearson's r | .060 | -.084 |
| | p-value | .527 | .370 |
| Race Diversity | Pearson's r | .035 | .207 |
| | p-value | .712 | ***.025** |
| Disability Diversity | Pearson's r | -.005 | .196 |
| | p-value | .959 | ***.034** |
| Sexual Orientation Diversity | Pearson's r | -.043 | .043 |
| | p-value | .645 | .643 |

*Note: *Significance level .05 is reached*

# 9. CONTRIBUTION & CONCLUSION

This dissertation aims to provide a new perspective on the topic of Biased AI by combining different fields of research such as Psychology, Law, Social Science, Business, and Computer Science. A comprehensive overview is developed structuring the issue down to its core components.

Primary, the fundamentals of AI, Bias and Discrimination are elaborated. Thereafter, the main type of biases and areas at risk for biased AI are identified namely racial bias, gender bias, age bias, sexual orientation bias, and disability bias. The main findings show that even though legal restrictions are imposed, human biases in these forms are prevalent if not inevitable.

Four main areas where biases in the AI system are most prevalent and extensively discussed in research are elaborated; employment sector, finance sector, criminal defense, and healthcare sector.

With respect to the employment sector, biases in hiring, firing and compensation are analyzed. Companies implementing AI systems in their hiring processes to increase efficiency and to be potentially more objective, however their systems lack transparency, accountability and fairness.

In the financial sector, half of the enterprises are investing heavily in AI systems (PwC study, 2017). These companies collect a large range of customer data as card payments are most prevalent and create detailed customer profiles. These data points can be used to make future decisions thereupon as for instance, for the individual assessment of granting loans and mortgages. As sensitive information is used, these data-driven decisions can be vulnerable to biases.

Regarding the criminal sector predictive policing is discussed whereby hotspots of potential crimes are predicted. Especially in the U.S., African American outnumber white people at alarming rates which numbers can be reinforced by biased algorithms. Automated facial recognition systems gain in escalating importance in police organizations around the world. Thereby, the “other-race effect” results in lower accuracy rates for minority groups that are therefore at risk of being wrongly convicted. Besides, predicting sentence length also poses a high risk for minorities as correlation and causation is not distinguished and therefore poses the risk in wrong presumptions of re-offending rates.

In the healthcare sector, the usage of AI systems contributes to today's medical advances, however, pose a risk of discrimination as for instance in the lower accuracy scores in skin cancer protection for dark-skinned patients.

The next research question aims to identify the major reasons for biased AI. Therefore, the development chain and the key stakeholders of an AI model are analyzed, and four main reasons are recognized; the used data, the responsible coder, the AI model itself and the model deployment.

First, the data is identified as one of the main reasons as it encodes past decisions made by biased humans. Further, a sample size disparity meaning an unbalanced representation of groups in the training data can lead to discrepancies in accuracy rates. This holds true for example for image recognition tools whereby models are mostly trained with white male images (Lohr, 2018; Buolamwini & Gebru, 2018). Further, the Simpson’s paradox can be related to fairness as distinct protective subgroups can be hidden within a larger subgroup and therefore discrimination can be

concealed in the datasets (Dwork et al., 2012). Additionally, biased AI can result from an outdated perception and decisions encoded in outdated data wherefore data has to be checked for validity.

Second, the coder creating a biased algorithm is analyzed as a potential reason. Statistics show that the creator of algorithms are mostly white men and might not have the comprehension or sympathy for creating fair AI models.

Third, the model is identified as a potential source of bias. Algorithms can be vulnerable towards bias as the system reinforces itself and lacks critical feedback. In the criminal sector this can mean that a predictive analytic tool targets neighborhoods with a disproportionate number of black citizens based on past crimes and therefore, increases the likelihood of crime committed by black individuals rather than by white individuals. Another risk is that AI models are in many cases black boxes lacking transparency and comprehensibility. Especially, the increased usage of deep learning algorithms produces results that are in many instances not retractable and revisable for their creators nor victims. Next, the new concept of a should-be model is briefly discussed. Usually, AI models translate outdated behavior patterns into the future and should be revised to include new ethical understandings, new social standards to promote contemporary, fair and social decisions. In addition, the risk of causations is introduced as with an abundance of data points meaningless correlations can appear which will be wrongly interpreted as causations.

Fourth, the risk of model deployment is briefly discussed as it is the last instance of the final decision. As AI models are not flawless, final human judgement is oftentimes necessary. Therefore, in some cases it is crucial to consider the AI system's decisions as a suggestion rather than a call to action.

The first part of the solutions for biased AI aims to provide guidance on how to measure fairness and biases in AI systems. Table 2 is created conflating and critically examining primary research findings of assessment techniques such as general statistical fairness measures, definitions based predicted and actual outcomes, predicted probabilities and actual outcomes, statistical outcomes, similarity-based measures, causal reasoning and absolute measures.

The second part discusses on how to mitigate biases and is subdivided into three parts: pre-, in- and post-processing. In-Processing focuses on altering the training data so that a model can be trained with a fair data set to produce fair results. In-processing focuses on identifying a fair classifier to result in algorithmic fairness without modifying the training data itself. A post-processing approach is applied if the training data or the learning algorithm cannot be modified (Bellamy et al., 2018). In contrast, a fair post-processor is learned from a predicted dataset and concluded in a fair predicted data set.

To evince the limitations of fairness measures and mitigation techniques, a comparison of the volume of research findings in the field of supervised and unsupervised learning is conducted. The opposition has shown that supervised learning has been more thoroughly researched with roughly double the research findings available for various search terms than for unsupervised learning (Table 5). A similar discrepancy can be found in the comparison between classification and regression as fairness in classification is more widely researched than in regressions. The discrepancy does, however, not hold true for its risk evaluation as discrimination can also arise in unsupervised learning and regressions. One additional limitation is that most prominent fairness measures and mitigation techniques are tested with one particular classification data set of a German credit card company. To thoroughly analyze and profoundly evidence fairness assessment mitigation measures, it is crucial to use high bandwidth of data sets representing different real-life challenges.

As part of this thesis, a survey was conducted registering the assessed risk of AI systems in comparison to human decision-making. Especially, their perception with respect to biased AI training and group diversity is analyzed. Thereupon, three hypotheses are developed and statistically tested. Hypothesis 1 was proven to be valid that individuals perceive AI systems as fairer and less biased than humans. However, the previous research conducted in this dissertation has proven that AI systems pose a real risk and have to be considered with caution. Survey respondents can underestimate that risk as they are not thoroughly educated on the issue on biased AI. Further, Hypothesis 2 analyzes the impact of training on Biased AI / algorithmic fairness on higher awareness for the risk of Biased AI systems which did not result in significant findings. In Hypothesis 3 the influence of team diversity of respondents in the risk assessment has shown that race and disability diversity correlates significantly positively with the risk score.

To conclude, in this thesis, we do a literature review, by putting in context much of the existent research on AI biases, and by structuring its core components. This analysis is complemented by new findings resulting from the quantitative analysis of the survey, which we hope will raise awareness, inform, and provide understanding on how to identify and mitigate biases, to the broader public, to companies, and to anyone interested in the topic of AI.

## 10.BIBLIOGRAPHY


Accenture (2017). Embracing artificial intelligence. Retrieved from https://www.accenture.com/_acnmedia/accenture/next-gen-5/event-g20-yea-summit/pdfs/accenture-intelligent-economy.pdf

Adebayo, J., & Kagal, L. (2016). Iterative orthogonal feature projection for diagnosing bias in black-box models. *arXiv preprint arXiv:1611.04967*.

Agarwal, S., Haq, R., & Coutinho, R. (2019, September 11). Fair AI: How to detect and remove bias from financial services AI models. *Finextra.* Retrieved from https://www.finextra.com/blogposting/17864/fair-ai-how-to-detect-and-remove-bias-from-financial-services-ai-models

Agence France-Presse. (2019, September 4). Smile-to-pay: Chinese shoppers turn to facial payment technology. *The Guardian*. Retrieved from https://www.theguardian.com/world/2019/sep/04/smile-to-pay-chinese-shoppers-turn-to-facial-payment-technology

AI Now Institute. (n.d.). A research institute examining the social implications of artificial intelligence. Retrieved from https://ainowinstitute.org/

Aiyar, M. S., & Ebeke, M. C. H. (2017). *The impact of workforce aging on European productivity*. International Monetary Fund.

Ajunwa, I. (2019, October 8). Beware of Automated Hiring. *The New York Times.* Retrieved from https://www.nytimes.com/2019/10/08/opinion/ai-hiring-discrimination.html

Ajunwa, I., Friedler, S., Scheidegger, C. E., & Venkatasubramanian, S. (2016). Hiring by algorithm: Predicting and preventing disparate impact. *Available at SSRN*.

Alain, G., & Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Alpaca (n.d.). Retrieved January 5, 2020 from https://www.alpaca.ai/en_forecast.html

Ameri, M., Schur, L., Adya, M., Bentley, F. S., McKay, P., & Kruse, D. (2018). The disability employment puzzle: A field experiment on employer hiring behavior. *ILR Review*, *71*(2), 329-364.

American Anthropological Association. (1998). AAA statement on race. Retrieved January 10, 2020 from https://www.americananthro.org/ConnectWithAAA/Content.aspx?ItemNumber=2583

Amnesty International. (2018). Trapped in the Matrix: Secrecy, stigma, and bias in the Met's gangs database.

Anyoha, R. (2017). The history of artificial intelligence. *Science in the News*, *28*.

Askham, N., Cook, D., Doyle, M., Fereday, H., Gibson, M., Landbeck, U., ... & Schwarzenbach, J. (2013). The Six primary dimension for data quality assessment. *DAMA United Kingdom*. Retrieved from https://www.whitepapers.em360tech.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf

Ayasdi (2018). Anti-money laundering solution deep dive. *Amazon AWS*. Retrieved from https://s3.amazonaws.com/cdn.ayasdi.com/wp-content/uploads/2018/04/22170635/AML_Solutions_Deep_Dive_WP_051617v01.pdf

Bachinskiy, A. (2019, February 21). The growing impact of AI in financial services: Six examples. *Towards Data Science.* Retrieved from https://towardsdatascience.com/the-growing-impact-of-ai-in-financial-services-six-examples-da386c0301b2

Badgett, M. L. (1995). The wage effects of sexual orientation discrimination. *ILR Review*, *48*(4), 726-739.

Badgett, L., & Frank, J. (Eds.). (2007). Sexual orientation discrimination: An international perspective. Routledge.

Badgett, M. V., & Schneebaum, A. (2015). The impact of wage equality on sexual orientation poverty gaps.

Bai, Y., & Ghanem, B. (2017). Multi-scale fully convolutional network for face detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 132-141).

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*, 671.

Baron, J. (2000). *Thinking and deciding*. Cambridge University Press.

Baumeister, R. F., & Leary, M. R. (1997). Writing narrative literature reviews. *Review of general psychology*, *1*(3), 311-320.

Beck, T., Behr, P., & Madestam, A. (2018). Sex and credit: Is there a gender bias in lending?. *Journal of Banking and Finance*, *87*.

Bell, M. (2017). Data collection in relation to LGBTI people. *Brüssel: European Commission*.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Nagar, S. (2018). AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 0049124118782533.

Beutel, A., Chen, J., Zhao, Z., & Chi, E. H. (2017). Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*.

Bickle, G. S., & Peterson, R. D. (1991). The impact of gender-based family roles on criminal sentencing. *Social Problems*, *38*(3), 372-394.

Binns, R. (2020, January). On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 514-524).

Black, D. A., Makar, H. R., Sanders, S. G., & Taylor, L. J. (2003). The earnings effects of sexual orientation. *ILR Review*, *56*(3), 449-469.

Black, H., Schweitzer, R. L., & Mandell, L. (1978). Discrimination in mortgage lending. *The American Economic Review*, *68*(2), 186-191.

Blandford, J. M. (2000). Evidence of the role of sexual orientation in the determination of earnings outcomes. *University of Chicago*.

Bloom v. AC T,Inc,. 18-cv-06749 (2018).

Bobo, L. D. (2001). Racial attitudes and relations at the close of the twentieth century. *America becoming: Racial trends and their consequences*, *2*, 264.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems* (pp. 4349-4357).

Book, E. W., & Book, E. W. (2000). *Why the best man for the job is a woman: The unique female qualities of leadership*. New York, NY: HarperBusiness.

Bouch, D. C., & Thompson, J. P. (2008). Severity scoring systems in the critically ill. *Continuing education in anaesthesia, critical care & pain*, *8*(5), 181-185.

Brault, M. (2008). Americans with disabilities: 2005. Current population reports, P70-117, US Census Bureau, Washington, DC.

Brayne, S. (2017). Big data surveillance: The case of policing. *American sociological review*, *82*(5), 977-1008.

Brayne, S., Rosenblat, A., & Boyd, D. (2015). Predictive policing. *Data & Civil Rights: A New Era Of Policing And Justice*.

Buchmueller, T., & Carpenter, C. S. (2010). Disparities in health insurance coverage, access, and outcomes for individuals in same-sex versus different-sex relationships, 2000–2007. *American journal of public health*, *100*(3), 489-495.

Bughin, J., Seong, J., Manyika, J., Chui, M., & Joshi, R. (2018). Notes from the AI frontier: Modeling the impact of AI on the world economy. *McKinsey Global Institute*.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91).

Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, *21*(2), 277-292.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186.

Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems* (pp. 3992-4001).

Calude, C. S., & Longo, G. (2017). The deluge of spurious correlations in big data. *Foundations of science*, *22*(3), 595-612.

Cambridge Dictionary (n.d.-a). Bias. In *Cambridge Dictionary*. Retrieved January 15, 2020 from https://dictionary.cambridge.org/dictionary/english/bias

Cambridge Dictionary. (n.d.-b). Discrimination. In *Cambridge Dictionary.* Retrieved February 10, 2020, from https://dictionary.cambridge.org/dictionary/english/discrimination

Career Builder. (2017, May 18). *More Than Half of HR Managers Say Artificial Intelligence Will Become a Regular Part of HR in Next 5 Years* [Press release]. Retrieved from http://press.careerbuilder.com/2017-05-18-More-Than-Half-of-HR-Managers-Say-Artificial-Intelligence-Will-Become-a-Regular-Part-of-HR-in-Next-5-Years

Carson, E. A. (2018). Prisoners in 2016 (No. NCJ 251149). US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

Casacuberta, D. (2018, May 9). Bias in a feedback loop: Fuelling algorithmic injustice. *CCCB Lab.* Retrieved from http://lab.cccb.org/en/bias-in-a-feedback-loop-fuelling-algorithmic-injustice/

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. *BMJ Qual Saf*, *28*(3), 231-237.

Chamorro-Premuzic, T. (2019, June 10). Will AI reduce gender bias in hiring?. Retrieved from https://hbr.org/2019/06/will-ai-reduce-gender-bias-in-hiring

Champion, D. J. (1987). Elderly felons and sentencing severity: Interregional variations in leniency and sentencing trends. *Criminal Justice Review*, *12*(2), 7-14.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of abnormal psychology*, *74*(3), 271.

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. *The New England journal of medicine*, *378*(11), 981.

Charlton, E. (2019, October 1). 6 things to know about China's historic rise. *World Economic Forum. R*etrieved from https://www.weforum.org/agenda/2019/10/china-economy-anniversary/

Charlton, B. M., Gordon, A. R., Reisner, S. L., Sarda, V., Samnaliev, M., & Austin, S. B. (2018). Sexual orientation-related disparities in employment, health insurance, healthcare access and health-related quality of life: a cohort study of US male and female adolescents and young adults. *BMJ open*, *8*(6), e020418.

Cheatham, B., Javanmardian, K., & Samandari, H. (2019). Confronting the risks of artificial intelligence. *McKinsey Quarterly.*

Chen, I., Johansson, F. D., & Sontag, D. (2018). Why is my classifier discriminatory?. In *Advances in Neural Information Processing Systems* (pp. 3539-3550).

Cheung, H. (2020, June 8). George Floyd death: Why US protests are so powerful this time. *BBC News*. Retrieved from https://www.bbc.com/news/world-us-canada-52969905

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, *5*(2), 153-163.

Chui, M., & Malhotra, S. (2018). AI adoption advances, but foundational barriers remain. *McKinsey and Company*.

Clain, S. H., & Leppel, K. (2001). An investigation into sexual orientation discrimination as an explanation for wage differences. *Applied economics*, *33*(1), 37-47.

Collier, L. (2014). Incarceration nation.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017, August). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 797-806).

Cossins, D. (2018). Discriminating algorithms: 5 times AI showed prejudice. *New Scientist*, *12*.

Couchman, H. (2019). Policing by machine: Predictive policing and the threat to our rights.

Craft, J. A., Doctors, S. I., Shkop, Y. M., & Benecki, T. J. (1979). Simulated management perceptions, hiring decisions and age. *Aging and Work*, *2*(2), 95-102.

Crew, B. (2019, August 2). Google Scholar reveals its most influential papers for 2019. *Nature Index*. Retrieved June 21, 2020 from https://www.natureindex.com/news-blog/google-scholar-reveals-most-influential-papers-research-citations-twenty-nineteen#:~:text=Tracking%20citation%20information%20for%20almost,and%20influence%22%20of%20recent%20publications.

Cronin, P., Ryan, F., & Coughlan, M. (2008). Undertaking a literature review: a step-by-step approach. *British journal of nursing*, *17*(1), 38-43.

Cuberes, D., & Teignier, M. (2016). Aggregate effects of gender gaps in the labor market: A quantitative estimate. *Journal of Human Capital*, *10*(1), 1-32.

Cutshall, C. R., & Adams, K. (1983). Responding to older offenders: Age selectivity in the processing of shoplifters. *Criminal Justice Review*, *8*(2), 1-8.

Daly, K., & Bordt, R. L. (1995). Sex effects and sentencing: An analysis of the statistical literature. *Justice Quarterly*, *12*(1), 141-175.

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

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *Proceedings on privacy enhancing technologies*, *2015*(1), 92-112.

Davison, H. K., & Burke, M. J. (2000). Sex discrimination in simulated employment contexts: A meta-analytic investigation. *Journal of Vocational Behavior*, *56*(2), 225-248.

Demirguc-Kunt, A., Klapper, L., & Singer, D. (2013). *Financial inclusion and legal discrimination against women: evidence from developing countries*. The World Bank.

Derrick, B., Toher, D., & White, P. (2017). How to compare the means of two samples that include paired observations and independent observations: A companion to Derrick, Russ, Toher and White (2017). *The Quantitative Methods in Psychology*, *13*(2).

Derrick, B., Russ, B., Toher, D., & White, P. (2017). Test statistics for the comparison of means for two samples that include both paired and independent observations. *Journal of Modern Applied Statistical Methods*, *16*(1), 9.

De Shutter, O., Lemmens, P., Kukova, S., Sturma, P., Olsen, B. K., Bielefeldt, H., ... & O'Connel, D. (2009). *Homophobia and discrimination on grounds of sexual orientation in the EU Member States. Part I-Legal analysis*. Publications Office of the European Union.

Dickinson, H. & Kavanagh, A. (2020, March 26). People with a disability are more likely to die from coronavirus – but we can reduce this risk. *The Conversation*. Retrieved from https://theconversation.com/people-with-a-disability-are-more-likely-to-die-from-coronavirus-but-we-can-reduce-this-risk-134383

Dovidio, J. F., & Gaertner, S. L. (2004). Aversive racism. *Advances in experimental social psychology*, *36*, 4-56.

Dreyfus, H. L., & Dreyfus, S. E. (1992). What artificial experts can and cannot do. *AI & society*, *6*(1), 18-26.

Drydakis, N. (2009). Sexual orientation discrimination in the labour market. *Labour Economics, 16*(4), 364-372.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214-226).

Eagly, A. H. (2013). *Sex differences in social behavior: A social-role interpretation*. Psychology Press.

Eagly, A. H., & Johannesen-Schmidt, M. C. (2001). The leadership styles of women and men. *Journal of social issues*, *57*(4), 781-797.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological review*, *109*(3), 573.

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American psychologist*, *49*(8), 709.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115-118.

European Data Protection Supervisor. (2018). Police directive. *European Data Protection Supervisor.* Retrieved from https://edps.europa.eu/data-protection/our-work/subjects/police-directive_en

European Union Agency for Fundamental Rights. (n.d.). EU LGBT survey technical report. *European Union Agency for Fundamental Rights*. Retrieved from https://fra.europa.eu/sites/default/files/eu-lgbt-survey-technical-report_en.pdf

European Union Agency for Fundamental Rights. (2009). How EU law offers protection from sexual orientation discrimination. Retrieved from https://fra.europa.eu/sites/default/files/fra_uploads/1227-Factsheet-homophobia-protection-law_EN.pdf

Eurostat. (2020, March 6). Women's employment in the EU. *Eurostat Your key European Statistics.* Retrieved March 5, 2020 from https://ec.europa.eu/eurostat/web/products-eurostat-news/-/EDN-20200306-1

Feagin, J. R. (1991). The continuing significance of race: Antiblack discrimination in public places. *American sociological review*, 101-116.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015, August). Certifying and removing disparate impact. InProceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 259-268). ACM.

Ferguson, A. G. (2019). *The rise of big data policing: Surveillance, race, and the future of law enforcement*. NYU Press.

Ferris, G. (2018). Face Off: The lawless growth of facial recognition in UK policing. *Big Brother Watch.* Retrieved from https://bigbrotherwatch.org.uk/wp-content/uploads/2018/05/Face-Off-final-digital-1.pdf

Ficarrotto, T. J. (1990). Racism, sexism, and erotophobia: Attitudes of heterosexuals toward homosexuals. *Journal of Homosexuality*, *19*(1), 111-116.

FinTech Futures. (2019, October 10). AI: Understanding bias and opportunities in financial services. *FinTech Futures.* Retrieved from https://www.fintechfutures.com/2019/10/ai-understanding-bias-and-opportunities-in-financial-services/

Finkelstein, L. M., & Burke, M. J. (1998). Age stereotyping at work: The role of rater and contextual factors on evaluations of job applicants. The Journal of general psychology, 125(4), 317-345.

Finkelstein, L. M., Burke, M. J., & Raju, M. S. (1995). Age discrimination in simulated employment contexts: An integrative analysis. Journal of applied psychology, 80(6), 652.

Flovik, V. (2019, August 14). The hidden risk of AI and big data. *Towards Data Science*. Retrieved from https://towardsdatascience.com/the-hidden-risk-of-ai-and-big-data-3332d77dfa6

Forbes Insights, & Intel AI (2019, March 27). Rethinking privacy for the AI era. *Forbes.* Retrieved from https://www.forbes.com/sites/insights-intelai/2019/03/27/rethinking-privacy-for-the-ai-era/#24dfbe3b7f0a

Fox, J. (2020, May 7). Coronavirus deaths by age: How it's like (and not like) other disease. Bloomberg. Retrieved June 5th from https://www.bloomberg.com/opinion/articles/2020-05-07/comparing-coronavirus-deaths-by-age-with-flu-driving-fatalities

Freedom for All Americans. (n.d). LGBTQ Americans aren't fully protected from discrimination in 30 states. Retrieved from https://www.freedomforallamericans.org/states/

Fridensköld, J. (2019). Biased AI: The hidden problem that needs an answer.

Furness, D. (2018, November 29). Why is AI today's "most important" technology? Ask Microsoft's Chef Envisioner. Retrieved from https://emerj.com/ai-future-outlook/why-is-ai-todays-most-important-technology/

Gajane, P., & Pechenizkiy, M. (2017). On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*.

Galhotra, S., Brun, Y., & Meliou, A. (2017, August). Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering* (pp. 498-510).

Garbade, D. M. J. (2018, September 4). Clearing the confusion: AI vs Machine Learning vs Deep Learning differences. *Towards Data Science*. Retrieved from https://towardsdatascience.com/clearing-the-confusion-ai-vs-machine-learning-vs-deep-learning-differences-fce69b21d5eb

Garvie, C., & Frankle, J. (2016). Facial-recognition software might have a racial bias problem. *The Atlantic*, *7*.

Garvie, C., Bedoya, A., & Frankle, J. (2016). The Perpetual Line-Up. Unregulated police face recognition in America. Georgetown Law Center on Privacy & Technology, October 18, 2016.

Gates, K. A. (2011). *Our biometric future: Facial recognition technology and the culture of surveillance* (Vol. 2). NYU Press.

Gergen K.J., & Gergen, M.M. (1985). Social psychology, New York: Springer.

Gibbs, J. T. (1988). Young, black, and male in America: An endangered species. Auburn House Publishing Company, 14 Dedham Street, Dover, MA 02030.

Google Scholar. (n.d.). About Google Scholar. Retrieved June1, 2020 from https://scholar.google.com/intl/en/scholar/about.html

Grigg, T. (2019, December 9). Simpson's paradox and interpreting data. *Towards Data Science.* Retrieved from https://towardsdatascience.com/simpsons-paradox-and-interpreting-data-6a0443516765

Guenole, N., & Feinzig, S. (2018). The business case for AI in HR. *With Insights and Tips on Getting Started. Armonk: IBM Smarter Workforce Institute, IBM Corporation.*

Hall, W., & Pesenti, J. (2017). Growing the artificial intelligence industry in the UK. *Department for Digital, Culture, Media & Sport and Department for Business, Energy & Industrial Strategy. Part of the Industrial Strategy UK and the Commonwealth*.

Han, H., & Jain, A. K. (2014). Age, gender and race estimation from unconstrained face images. *Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, MSU Tech. Rep. (MSU-CSE-14-5)*, *87*, 27.

Hao, K. (2019a, January 21). AI is sending people to jail—and getting it wrong. *MIT Technology Review*. Retrieved from https://www.technologyreview.com/s/612775/algorithms-criminal-justice-ai/

Hao, K. (2019b, December 20). A US government study confirms most face recognition systems are racist. *MIT Technology Review*. Retrieved from https://www.technologyreview.com/f/614986/ai-face-recognition-racist-us-government-nist-study/

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems* (pp. 3315-3323).

Hardyns, W., & Rummens, A. (2018). Predictive policing as a new tool for law enforcement? Recent developments and challenges. *European journal on criminal policy and research*, *24*(3), 201-218.

Hart, C. (2018). *Doing a literature review: Releasing the research imagination*. Sage.

Hartocollis, A. (2018, June 15). Applicants lower on personality traits, suit Says. Retrieved May 20, 2020 from https://www.nytimes.com/2018/06/15/us/harvard-asian-enrollment-applicants.html?auth=login-google

Heider, F. (1958). The psychology of interpersonal relations. New York: John Wiley.

Henderson-King, E. I., & Nisbett, R. E. (1996). Anti-Black prejudice as a function of exposure to the negative behavior of a single Black person. *Journal of Personality and Social Psychology*, *71*(4), 654.

Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., Minhas, R., Sheikh, A., & Brindle, P. (2008). Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *Bmj*, *336*(7659), 1475-1482.

Houser, K. (2019, April 17). Experts: To solve AI's bias problem, hire fewer white men. *Towards Data Science*. Retrieved from https://futurism.com/biased-ai-problem-hire-fewer-white-males

Housman, M., & Minor, D. (2015). Toxic workers. Harvard Business School Strategy Unit Working Paper, (16-057).

Howard, J. W., & Rothbart, M. (1980). Social categorization and memory for in-group and out-group behavior. *Journal of Personality and Social Psychology*, *38*(2), 301.

Human Rights Campaign. (2019, October 1). The equality act. Retrieved from https://www.hrc.org/resources/the-equality-act

IBM Corporation. (n.d.-a). AI bias will explode. But only the unbiased AI will survive. Retrieved December 20, 2019 from https://www.research.ibm.com/5-in-5/ai-and-bias/

IBM Corporation. (n.d.-b). IBM Talent Management Software. Retrieved January 20, 2020 from https://www.ibm.com/talent-management/ai-in-hr-business-case/#section-3

IBM Corporation. (2018). Cognitive Compensation: Driving smarter decisions with transformative technology. Retrieved from https://www.ibm.com/downloads/cas/74KAB2EQ

IBM Research. (2019). Five innovations that will help change our lives within five years. Retrieved from https://www.research.ibm.com/5-in-5/ai-and-bias/

INTERPOL (n.d.). Facial recognition. Retrieved January 26, 2020 from https://www.interpol.int/en/How-we-work/Forensics/Facial-Recognition

Jansen, F. (2018). Data Driven Policing in the Context of Europe. *Data Justice Lab*.

JMC Stanford (n.d.) What is AI? / basic questions. Retrieved from http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html

Jaschik, S. (2017, August 7). The Numbers and the arguments on Asian admissions. Retrieved May 5, 2020 from https://www.insidehighered.com/admissions/article/2017/08/07/look-data-and-arguments-about-asian-americans-and-admissions-elite

Jaschik, S. (2018, April 27). Making the case for test optional. Inside Higher Ed. Retrieved March 20, 2020 from https://www.insidehighered.com/news/2018/04/27/large-study-finds-colleges-go-test-optional-become-more-diverse-and-maintain

Kaebble, D., & Mary, C. (2018). Correctional populations in the United States, 2016. *US. Department of Justice.*

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological review*, *80*(4), 237.

Kahneman, D., Slovic, S. P., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.

Kamiran, F., & Calders, T. (2009, February). Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication* (pp. 1-6). IEEE.

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1-33.

Kamishima, T., Akaho, S., & Sakuma, J. (2011, December). Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops* (pp. 643-650). IEEE.

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012, September). Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 35-50). Springer, Berlin, Heidelberg.

Katsarou, M. (2019). Women & the leadership labyrinth: Howard vs Heidi. *Leadership Psychology Institute, Leadership Psychology Institute, www. leadershippsychologyinstitute. com/women-the-leadership-labyrinthhoward-vs-heidi*.

Keller, S., & Meaney, M. (2017). Attracting and retaining the right talent. *McKinsey.*

Khalid, A. (2019, October 16). Facial recognition AI can't identify trans and non-binary people. Quartz. Retrieved from https://qz.com/1726806/facial-recognition-ai-from-amazon-microsoft-and-ibm-misidentifies-trans-and-non-binary-people/

Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems* (pp. 656-666).

Klare, B. F., Burge, M. J., Klontz, J. C., Bruegge, R. W. V., & Jain, A. K. (2012). Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, *7*(6), 1789-1801.

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological review*, *94*(2), 211.

Kleck, G. (1981). Racial discrimination in criminal sentencing: A critical evaluation of the evidence with additional evidence on the death penalty. *American Sociological Review*, 783-805.

Klein, D., & Kress, J. (1976). Any woman's blues: A critical overview of women, crime and the criminal justice system. *Crime and Social Justice*, (5), 34-49.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Knight, W. (2017a, July 12). Biased algorithms are everywhere, and no one seems to care. *MIT Technology Review*. Retrieved from https://www.technologyreview.com/s/608248/biased-algorithms-are-everywhere-and-no-one-seems-to-care/

Knight, W. (2017b, October 3). Forget killer robots—bias is the real AI danger. *MIT Technology Review.* Retrieved from https://www.technologyreview.com/s/608986/forget-killer-robotsbias-is-the-real-ai-danger/

Koppel, R., Metlay, J. P., Cohen, A., Abaluck, B., Localio, A. R., Kimmel, S. E., & Strom, B. L. (2005). Role of computerized physician order entry systems in facilitating medication errors. *Jama*, *293*(10), 1197-1203.

Kozyrkov, C. (2019, January 24). What is AI bias?. Towards Data Science. Retrieved from https://towardsdatascience.com/what-is-ai-bias-6606a3bcb814

Kruglanski, A. W., & Ajzen, I. (1983). Bias and error in human judgment. *European Journal of Social Psychology*, *13*(1), 1-44.

Kruse, D., Schur, L., Rogers, S., & Ameri, M. (2018). Why do workers with disabilities earn less? Occupational job requirements and disability discrimination. *British Journal of Industrial Relations*, *56*(4), 798-834.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems* (pp. 4066-4076).

Langton, L., & Durose, M. R. (2013). *Police behavior during traffic and street stops, 2011*. Washington, DC: US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

Leetaru, K. (2019, January 15). A reminder that machine learning is about correlations not causation. *Forbes*. Retrieved from https://www.forbes.com/sites/kalevleetaru/2019/01/15/a-reminder-that-machine-learning-is-about-correlations-not-causation/#6e8bdba06161

Levi, G., & Hassner, T. (2015). Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 34-42).

Levin, W. C. (1988). Age stereotyping: College student evaluations. *Research on Aging*, *10*(1), 134-148.

Lichman, M. (2013). UCI Machine Learning Repository. Retrieved from http://archive.ics. uci.edu/ml

Locke, C. (2019, July 5). Why gender bias still occurs and what we can do about it. *Forbes*. Retrieved from https://www.forbes.com/sites/londonschoolofeconomics/2019/07/05/why-gender-bias-still-occurs-and-what-we-can-do-about-it/#32fa56d65228

Loftus, J. R., Russell, C., Kusner, M. J., & Silva, R. (2018). Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*.

Lohr, S. (2018, February 9). Facial recognition is accurate, if you're a white guy. *The New York Times*. Retrieved from https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html

London Policing Ethics Panel (2018). Interim report on live facial recognition. Retrieved from http://www.policingethicspanel.london/uploads/4/4/0/7/44076193/lpep_report_-_live_facial_recognition.pdf

Mann, R. D. (1959). A review of the relationships between personality and performance in small groups. *Psychological bulletin*, 56(4), 241.

Maruti Techlabs. (n.d.). 5 Ways AI is transforming the finance industry. Retrieved January 20, 2020 from https://marutitech.com/ways-ai-transforming-finance/

Massachusetts Institute of Technology (2020). MIT facts enrollment 2019-2020. Retrieved from http://web.mit.edu/facts/enrollment.html

Massey, D. S. (2001). Residential segregation and neighborhood conditions in US metropolitan areas. *America becoming: Racial trends and their consequences*, *1*(1), 391-434.

Mathias, M., Benenson, R., Pedersoli, M., & Van Gool, L. (2014, September). Face detection without bells and whistles. In *European conference on computer vision* (pp. 720-735). Springer, Cham.

Mauer, M. (2011). Addressing racial disparities in incarceration. *The Prison Journal*, *91*(3_suppl), 87S-101S.

Mayor of London. (2018. December 21). Review of the MPS gangs matrix. *Greater London Authority.* Retrieved from https://www.london.gov.uk/mopac-publications-0/review-mps-gangs-matrix

McGuire, W. J. (1960). A syllogistic analysis of cognitive relationships. *Attitude organization and change*, 65-111.

Melloni, C., Berger, J. S., Wang, T. Y., Gunes, F., Stebbins, A., Pieper, K. S., ... & Newby, L. K. (2010). Representation of women in randomized clinical trials of cardiovascular disease prevention. *Circulation: Cardiovascular Quality and Outcomes*, *3*(2), 135-142.

Metz, C. (2018, April 19). A.I. researchers are making more than $1 million, even at a nonprofit. *The New York Times*. Retrieved from https://www.nytimes.com/2018/04/19/technology/artificial-intelligence-salaries-openai.html

Mirza, S. A., & Rooney, C. (2018). Discrimination prevents LGBTQ people from accessing health care. *Center for American Progress*, *18*.

Monroy, M. (2018, August 15). G20 in Hamburg: Data protection commissioner considers face recognition illegal. *Digit*. Retrieved from https://digit.site36.net/2018/08/15/g20-in-hamburg-data-protection-commissioner-considers-face-recognition-illegal/

Morrison, M. A., & Morrison, T. G. (2003). Development and validation of a scale measuring modern prejudice toward gay men and lesbian women. *Journal of homosexuality, 43*(2), 15-37.

Morrison, M. A., & Morrison, T. G. (2011). Sexual orientation bias toward gay men and lesbian women: Modern homonegative attitudes and their association with discriminatory behavioral intentions 1. *Journal of Applied Social Psychology, 4*1(11), 2573-2599.

Morrison, M. A., Morrison, T. G., & Franklin, R. (2009). Modern and old-fashioned homonegativity among samples of Canadian and American university students. *Journal of Cross-Cultural Psychology*, *40*(4), 523-542.

Nabi, R., & Shpitser, I. (2018, April). Fair inference on outcomes. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Nadler, J. T., & Kufahl, K. M. (2014). Marital status, gender, and sexual orientation: Implications for employment hiring decisions. *Psychology of Sexual Orientation and Gender Diversity, 1*(3), 270.

Narain, S. (2009). Access to finance for women SME entrepreneurs in Bangladesh. *World Bank Working Paper*.

Needham, J. (1976). *Science and civilisation in China* (Vol. 5). Cambridge University Press.

Nisbett, R. E., & Ross, L. (1980). Human inference: Strategies and shortcomings of social judgment.

Norris, P. (1987). Politics and sexual equality the comparative position of women in Western democracies.

Nurek, M., Kostopoulou, O., Delaney, B. C., & Esmail, A. (2015). Reducing diagnostic errors in primary care. A systematic meta-review of computerized diagnostic decision support systems by the LINNEAUS collaboration on patient safety in primary care. *European Journal of General Practice*, *21*(sup1), 8-13.

Olfat, M., & Aswani, A. (2019, July). Convex formulations for fair principal component analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 663-670).

Olszewska, J. I. (2016). Automated face recognition: challenges and solutions. *Pattern Recognition-Analysis and Applications*, 59-79.

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

Osoba, O. A., & Welser IV, W. (2017). *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Rand Corporation.

Oxford Learner's Dictionaries. (n.d.). Xenophobia. Oxford Learner's Dictionaries. Retrieved January 22, 2020 from https://www.oxfordlearnersdictionaries.com/definition/american_english/xenophobia

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition.

Paul, K. (2019, April 17). 'Disastrous' lack of diversity in AI industry perpetuates bias, study finds. *The Guardian.* Retrieved from https://www.theguardian.com/technology/2019/apr/16/artificial-intelligence-lack-diversity-new-york-university-study

Pearl, J. (2014). Comment: understanding Simpson's paradox. *The American Statistician*, *68*(1), 8-13.

Petersilia, J. (1983). *Racial disparities in the criminal justice system* (Vol. 2947). Santa Monica, CA: Rand Corporation.

Pierce, L., & Balasubramanian, P. (2015). Behavioral field evidence on psychological and social factors in dishonesty and misconduct. *Current Opinion in Psychology*, *6*, 70-76.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. In *Advances in Neural Information Processing Systems* (pp. 5680-5689).

Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. Nature News, 538(7624), 161.

PredPol. (n.d.). what. where. when. Retrieved December 19, 2019 from https://www.predpol.com/

Provost, F. (2000). Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI'2000 workshop on imbalanced data sets* (Vol. 68, pp. 1-3). AAAI Press.

Provost, F., & Kohavi, R. (1998). On applied research in machine learning. *MACHINE LEARNING-BOSTON-*, *30*, 127-132.

Puzzanchera, C., Sladky, A., & Kang, W. (2016). Easy access to juvenile populations: 1990–2015. *Retrieved April*, *19*, 2018.

PwC. (n.d.). Bot.me: A revolutionary partnership. Retrieved from http://pwcartificialintelligence.com/

PwC. (2017). Top financial services issues of 2018. Retrieved from https://www.pwc.com/us/en/financial-services/research-institute/assets/pwc-fsi-top-issues-2018.pdf

PwC. (2018). Women in Work Index 2018. Retrieved from https://www.pwc.co.uk/services/economics-policy/insights/women-in-work-index.html

Ragins, B. R., & Cornwell, J. M. (2001). Pink triangles: antecedents and consequences of perceived workplace discrimination against gay and lesbian employees. *Journal of applied psychology*, *86*(6), 1244.

Rao, A., & Verweij, G. (2017). Sizing the prize: What's the real value of AI for your business and how can you capitalise. *PwC Publication, PwC*.

Reiss, B. F., Safer, J., & Yotive, W. (1976). Psychological test data on female homosexuality: a review of the literature. *Journal of Homosexuality*, *1*(1), 71-86.

Phillips, P. J., Jiang, F., Narvekar, A., Ayyad, J., & O'Toole, A. J. (2011). An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception (TAP)*, *8*(2), 1-11.

Rhodes, S. R. (1983). Age-related differences in work attitudes and behavior: A review and conceptual analysis. *Psychological bulletin*, *93*(2), 328.

Ridley, M. (1996). The origins of virtue. London: Viking. *Penquin Books, P. Slovic (1972). Psychological study of human judgment: implications for investment decision making. Journal of Finance*, *27*(4), 779-799.

Rosenbluth, F., Salmond, R., & Thies, M. F. (2006). Welfare works: explaining female legislative representation. *Politics & Gender*, *2*(2), 165-192.

Ross, L., Greene, D., & House, P. (1977). The “false consensus effect”: An egocentric bias in social perception and attribution processes. *Journal of experimental social psychology*, *13*(3), 279-301.

Rothe, R., Timofte, R., & Van Gool, L. (2018). Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, *126*(2-4), 144-157.

Rudman, L. A., & Kilianski, S. E. (2000). Implicit and explicit attitudes toward female authority. *Personality and social psychology bulletin*, *26*(11), 1315-1328.

Rule, W. (1981). Why women don't run: The critical contextual factors in women's legislative recruitment. *Western Political Quarterly*, *34*(1), 60-77.

Salimi, B., Howe, B., & Suciu, D. (2019). Data Management for causal algorithmic fairness. *arXiv preprint arXiv:1908.07924*.

Salimi, B., Rodriguez, L., Howe, B., & Suciu, D. (2019, June). Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data* (pp. 793-810).

Sandberg, S. (2015). Lean in - Women, work and the will to lead.

SAS (n.d.). Artificial Intelligence: What it is and why it matters. Retrieved from https://www.sas.com/en_us/insights/analytics/what-is-artificial-intelligence.html

Sawyer, W., & Wagner, P. (2020). Mass incarceration: the whole pie 2020. *Prison Policy Initiative*.

Scherer, M. (2017). AI in HR: Civil rights implications of employer’s use of artificial intelligence and big data. Scitech Lawyer, 13(2), 12.

Schroer, A. (2019, May 23). AI and the bottom line: 15 examples of Artificial Intelligence in finance. *Builtin.* Retrieved from https://builtin.com/artificial-intelligence/ai-finance-banking-applications-companies

Scott, S. (2018). The war on gangs or a racialised war on working class black youths. *London: The Monitoring Group*.

Sellers, R. M., & Shelton, J. N. (2003). The role of racial identity in perceived racial discrimination. *Journal of personality and social psychology*, *84*(5), 1079.

Sentencing Project. (2018, April 19). Report to the United Nations on racial disparities in the U.S. criminal justice system. Retrieved March 10, 2020 from https://www.sentencingproject.org/publications/un-report-on-racial-disparities/

Shwartz-Ziv, R., & Tishby, N. (2017). Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*.

Silberg J. & Manyika, J. (June 6, 2019). Tackling bias in artificial intelligence (and in humans). McKinsey Global Institute.

Simoiu, C., Corbett-Davies, S., & Goel, S. (2017). The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, *11*(3), 1193-1216.

Smith, M. (2018, October 30). Can we predict when and where a crime will take place?. *BBC News.* Retrieved from https://www.bbc.com/news/business-46017239

Sobol, M. G., & Ellard, C. J. (1988). Measures of employment discrimination: A statistical alternative to the four-fifths rule. Industrial Relations Law Journal, 381-399.

Spohn, C. (1990). The sentencing decisions of black and white judges: Expected and unexpected similarities. *Law and Society Review*, 1197-1216.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research, 15*(1), 1929-1958.

Steffensmeier, D., Kramer, J., & Streifel, C. (1993). Gender and imprisonment decisions. *Criminology*, *31*(3), 411-446.

Steffensmeier, D., Kramer, J., & Ulmer, J. (1995). Age differences in sentencing. *Justice Quarterly*, *12*(3), 583-602.

Steffensmeier, D., Ulmer, J., & Kramer, J. (1998). The interaction of race, gender, and age in criminal sentencing: The punishment cost of being young, black, and male. *Criminology, 36*(4), 763-798.

Supreme Court of the United States. Griggs v. Duke Power Co. 401 U.S. 424, March 8, 1971.

Supreme Court of the United States. Ricci v. DeStefano. 557 U.S. 557, 174, 2009.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).

Tajfel, H. (Ed.). (2010). *Social identity and intergroup relations* (Vol. 7). Cambridge University Press.

The Economist. (2013, November 19). Discrimination abounds. Women are excluded from finance across the developing world. Retrieved March 20, 2019 from https://www.economist.com/free-exchange/2013/11/19/discrimination-abounds

Time Inc. (1970, November). Computers will be playing office politics. *Life Magazine, Vol. 69, No. 21.,* p.58.

Tobena, A., Marks, I., & Dar, R. (1999). Advantages of bias and prejudice: An exploration of their neurocognitive templates. *Neuroscience & Biobehavioral Reviews, 23*(7), 1047-1058.

Torraco, R. J. (2005). Writing integrative literature reviews: Guidelines and examples. *Human resource development review*, *4*(3), 356-367.

Total System Services Inc. (2019). Consumer payment study. Retrieved from https://www.tsys.com/Assets/TSYS/downloads/rs_2018-us-consumer-payment-study.pdf

Trewin, S. (2018). AI fairness for people with disabilities: Point of view. *arXiv preprint arXiv:1811.10670.*

Turing, I. B. A. (1950). Computing machinery and intelligence-AM Turing. *Mind*, *59*(236), 433.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, *5*(2), 207-232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124-1131.

Unicef. (2020). COVID-19 response: Considerations for children and adults with disabilities. Retrieved from https://www.unicef.org/disabilities/files/COVID-19_response_considerations_for_people_with_disabilities_190320.pdf

United Nations. (2007). Convention on the rights of persons with disabilities (CRPD). Un.org. Retrieved on March 10, 2020 from https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities/preamble.html

United Nations Women. (2018). Facts and figures: Economic empowerment. *Unicef*. Retrieved from https://www.unwomen.org/en/what-we-do/economic-empowerment/facts-and-figures

University of Sheffield. (n.d.). Protected characteristics. Retrieved January 10, 2020 from https://www.sheffield.ac.uk/hr/equality/focus/2.5491/protected

University of Southern California. (n.d.). Organizing your social sciences research paper. Retrieved March 20, 2020 from https://libguides.usc.edu/writingguide/literaturereview

U.S. department of health and human services office of minority health. (2019, August 22). Profile: Black/African Americans. Retrieved from https://www.minorityhealth.hhs.gov/omh/browse.aspx?lvl=3&lvlid=61

Van Brakel, R. (2016). Pre-emptive big data surveillance and its (dis) empowering consequences: the case of predictive policing. *pp. in*, 117-141.

Verma, S., & Rubin, J. (2018, May). Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)* (pp. 1-7). IEEE.

Weichselbaumer, D. (2003). Sexual orientation discrimination in hiring. *Labour Economics, 10*(6), 629-642.

Wen, Y., Zhang, K., Li, Z., & Qiao, Y. (2016, October). A discriminative feature learning approach for deep face recognition. In *European conference on computer vision* (pp. 499-515). Springer, Cham.

West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems: Gender, race and power in AI. *AI Now Institute*, 1-33.

Wilson, D. S., & Sober, E. (1994). Reintroducing group selection to the human behavioral sciences. *Behavioral and brain sciences*, *17*(4), 585-608.

Wilson, V., & Williams, J. (2019, September 11). Racial and ethnic income gaps persist amid uneven growth in household incomes. *Economic Policy Institute*. Retrieved from https://www.epi.org/blog/racial-and-ethnic-income-gaps-persist-amid-uneven-growth-in-household-incomes/

Wong, D. (2018, October 4). What exactly is an A.I. model? An ELi5. *Medium.* Retrieved from https://medium.com/datadriveninvestor/what-exactly-is-an-a-i-model-an-eli5-2b87e4d48114

World Bank Group. (2019). Women, business and the law 2019. *World Bank Publications*.

World Prison Brief. (2018). World prison brief data. *Prison Studies.* Retrieved from https://www.prisonstudies.org/country/united-states-america

Zaccaro, S. J., Kemp, C., & Bader, P. (2004). Leader traits and attributes. The nature of leadership, 101, 124.

Zafeiriou, S., Zhang, C., & Zhang, Z. (2015). A survey on face detection in the wild: past, present and future. *Computer Vision and Image Understanding*, *138*, 1-24.

Zehlike, M., Castillo, C., & Bonchi, F. (2017). Fairness measures: Datasets and software for detecting algorithmic discrimination. *Fairness-measures.* Retrieved from http://fairness-measures.org/

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013, February). Learning fair representations. In *International Conference on Machine Learning* (pp. 325-333).

Zhang, M. (2015, July 1). Google photos tags two African-Americans as gorillas through facial recognition software. *Forbes*. Retrieved from https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/#74161998713d

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018, December). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 335-340).

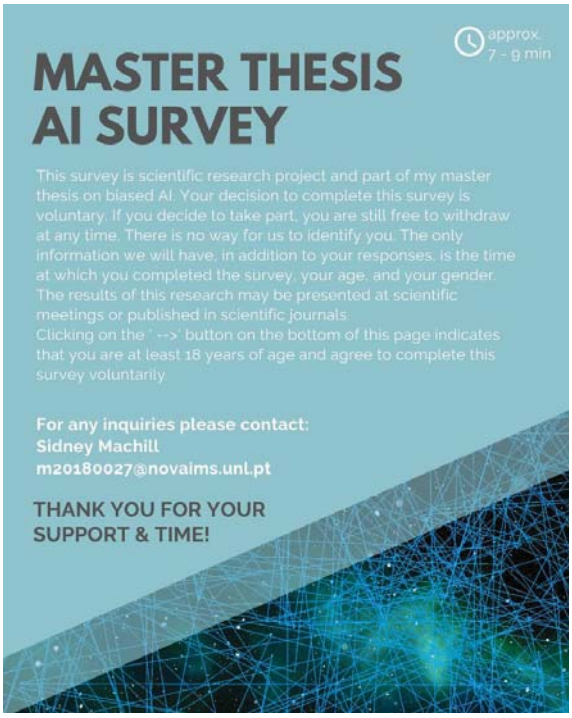
Zliobaite, I. (2015). On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*.

Zliobaite, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, *31*(4), 1060-1089.

# 11.APPENDIX

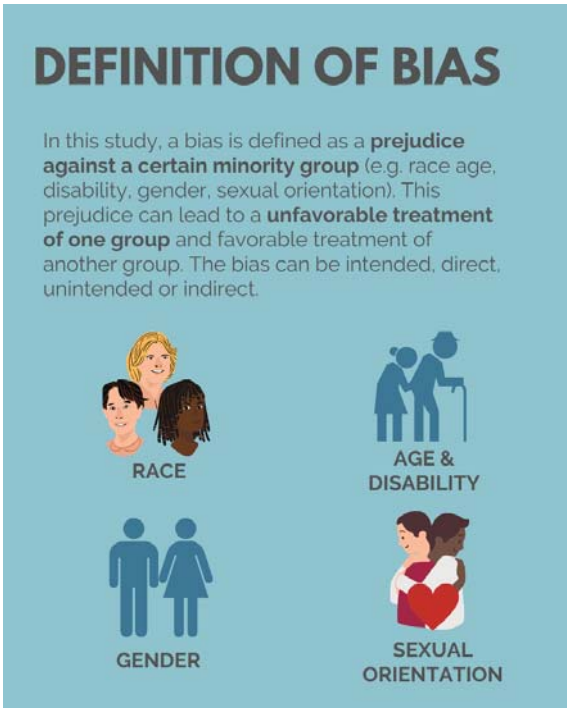
## Appendix A

*Welcome Message Survey*




## Appendix B

*Visualization "Definitions of Bias"*



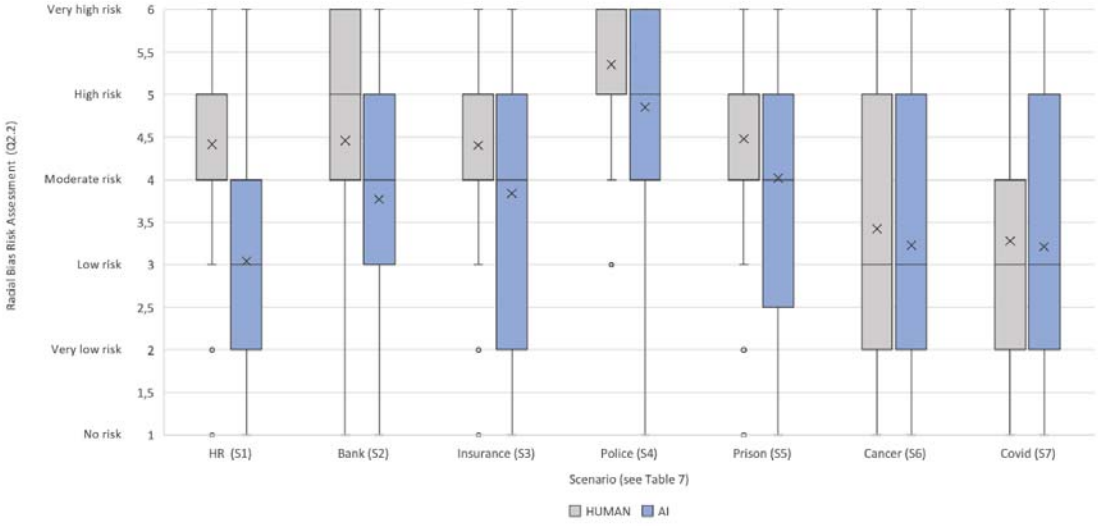
**Appendix C**

*Survy Results per Scenario*

| Scenario | Decision-Maker | Mean | Standard deviation |
|---|---|---|---|
| **S1 (HR)** | HU | 4.49 | .794 |
| | AI | 3.745 | 1.206 |
| **S2 (Bank)** | HU | 4.56 | 1.053 |
| | AI | 3.833 | 1.117 |
| **S3 (Insurance)** | HU | 4.596 | .869 |
| | AI | 4.023 | 1.267 |
| **S4 (Police)** | HU | 5.083 | .767 |
| | AI | 4.333 | 1.356 |
| **S5 (Prison)** | HU | 4.391 | 1.043 |
| | AI | 4.04 | 1.195 |
| **S6 (Cancer)** | HU | 3.708 | 1.148 |
| | AI | 3.333 | 1.155 |
| **S7 (Covid)** | HU | 4.06 | .956 |
| | AI | 4 | 1.272 |

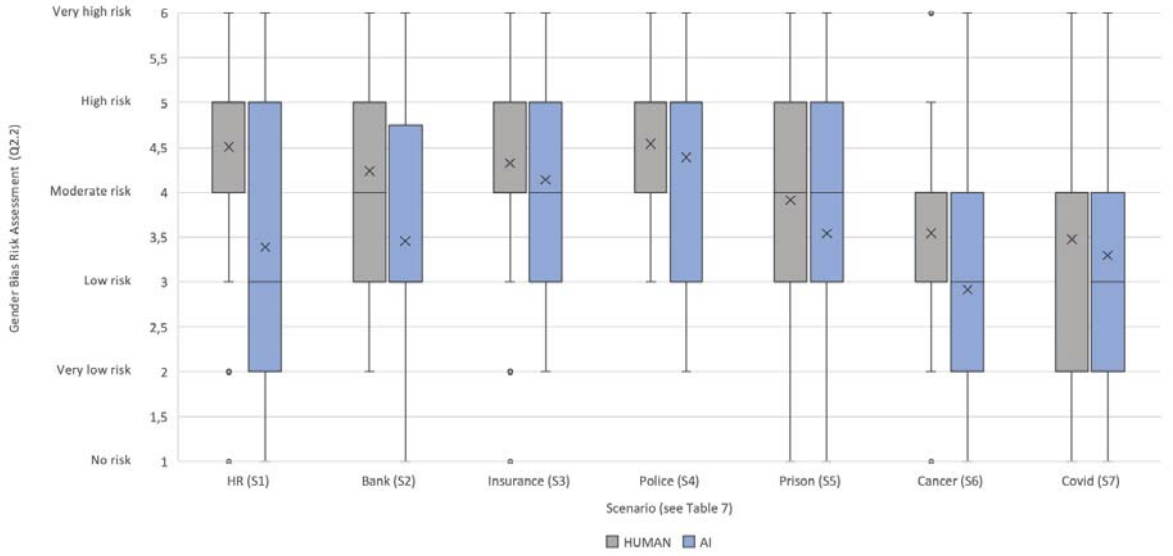
**Appendix D**

*Average Risk Assessment for Race Bias (Answer to Q2.2.)*



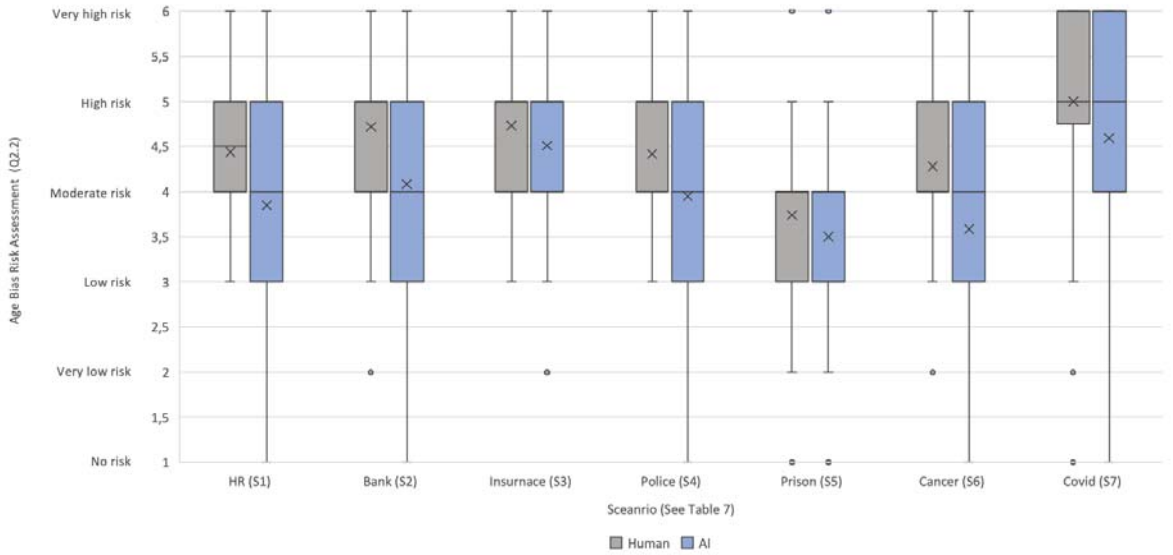
**Appendix E**

*Average Risk Assessment for Gender Bias (Answer to Q2.2.)*



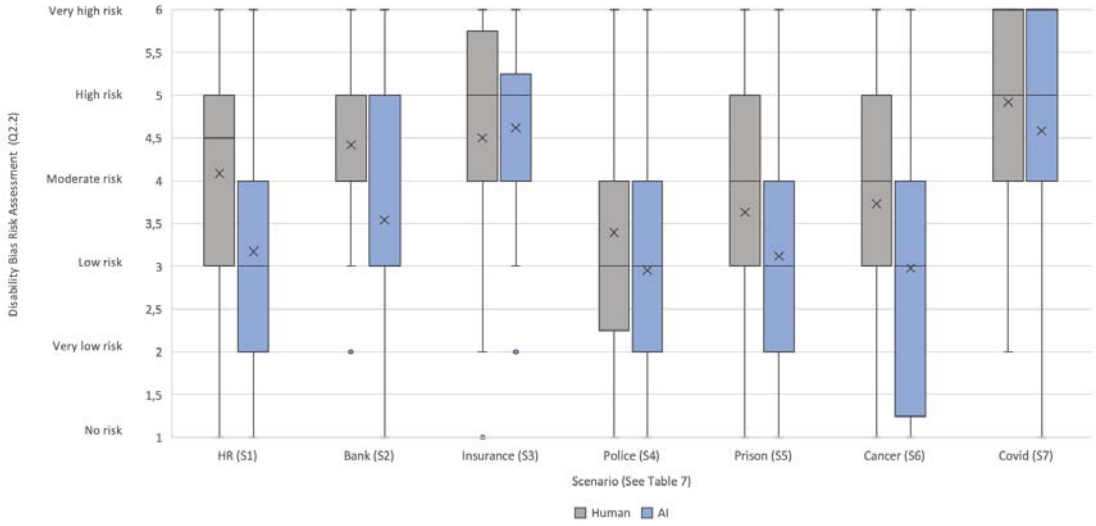

**Appendix F**

*Average Risk Assessment for Age Bias (Answer to Q2.2.)*



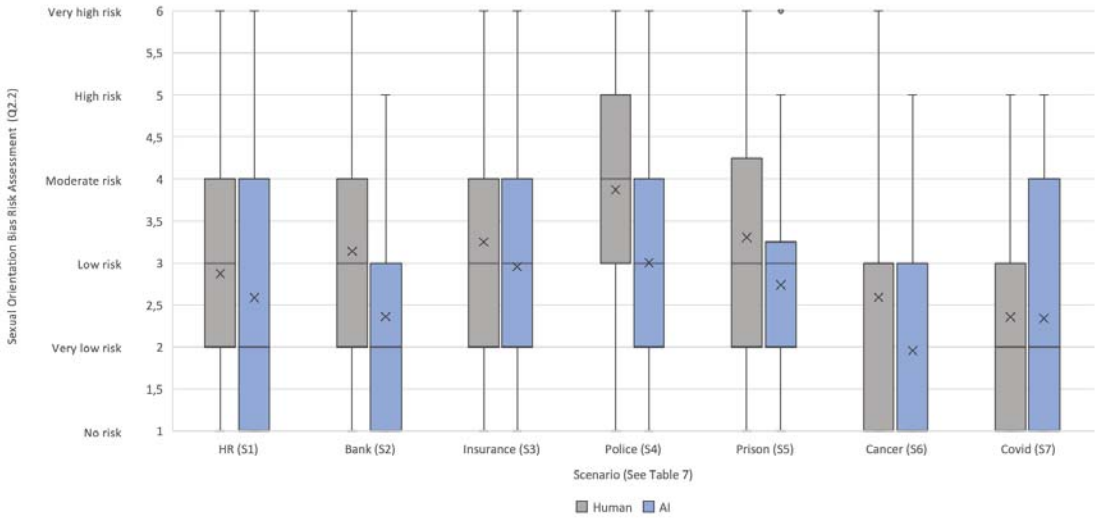
**Appendix G**

*Average Risk Assessment for Disability Bias (Answer to Q2.2.)*




**Appendix H**

*Average Risk Assessment for Sexual Orientation Bias (Answer to Q2.2.)*




**Appendix I**

*Independent Sample T-Test for Educational Background STEM and Uneuqal Business*

| Scenario | t | df | p |
|---|---|---|---|
| **HR_HU** | -0.910 | 46 | 0.368 |

| | | | |
|---|---|---|---|
| **HR_AI** | 0.035 | 41 | 0.972 |
| **Bank_HU** | -0.485 | 42 | 0.630 |
| **Bank_AI** | -1.758 | 39 | 0.087 |
| **Insurance_HU** | 1.657 | 47 | 0.104 |
| **Insurance_AI** | 0.104 | 36 | 0.918 |
| **Police_HU** | 2.364 | 38 | **0.023** |
| **Police_AI** | -0.144 | 38 | 0.886 |
| **Prison_HU** | -1.187 | 38 | 0.242 |
| **Prison_AI** | 0.453 | 41 | 0.653 |
| **Cancer_HU** | -0.643 | 41 | 0.524 |
| **Cancer_AI** | -0.918 | 43 | 0.364 |
| **Covid_HU** | 0.252 | 43 | 0.802 |
| **Covid_AI** | -0.135 | 42 | 0.893 |

*Note.* Student's t-test.
*Note.* STEM includes Data Science, Information Systems, Computer Science, Mathematics, and similar backgrounds
*Note.* Business includes management, finance, and similar backgrounds

**Appendix J**

*Paired t-Test*

Derrick, Russ, et al. (2017) defined the test statistic t as:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} - 2r\frac{s_1 s_2 n_c}{n_1 n_2}}}$$

Where $x_1$ mean is the average of all observations for sample 1 (here individuals' respondent to the AI scenario) and $x_2$ mean the average of all observations for sample 2 (here individuals responded to the human scenario). *s1* defines the standard deviation of the responses of sample 1 and $s_2$ the standard deviation of all observations of sample 2. Further, the variable *r* is the Pearson's correlation

coefficient for the paired observations solely (individuals who answered both scenarios). $n_a$ are the observations from sample 1, $n_b$ from sample 2 and $n_c$ are individuals belonging to both samples (here individuals answered both the scenario including an AI system and a human decision-maker). Further, the variables n1and n2are defined as $n_1=n_a+n_c$ and $n_2=n_b+n_c$.

The test statistic t follows a t-distribution, if the null hypothesis is true. The degree of freedom is defined as v:

$$v = (n_c - 1) + \frac{\gamma - n_c + 1}{n_a + n_b + 2n_c}(n_a + n_b)$$

where:

$$\gamma = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

**Appendix K**

*Code "partiallyoverlapping" in R*

The following code illustrates based on the example of the HR scenario (S1) the "partiallyoverlapping" package code in R.

```
install.packages('Partiallyoverlapping')
library(Partiallyoverlapping)

setwd("~/Downloads")

hu.paired=read.table("HR_HU_Paired.csv", sep=",", header=TRUE)
colnames(hu.paired)=c("id", "HR_hu_paired")

ai.paired=read.table("HR_AI_Paired.csv", sep=",", header=TRUE)
colnames(ai.paired)=c("id", "HR_ai_paired")

hu.unpaired=read.table("HR_HU_Unpaired.csv", sep=",", header=TRUE)
colnames(hu.unpaired)=c("id", "HR_hu_unpaired")

ai.unpaired=read.table("HR_AI_Unpaired.csv", sep=",", header=TRUE)
colnames(ai.unpaired)=c("id", "HR_ai_unpaired")

Partover.test(ai.unpaired$HR_ai_unpaired, hu.unpaired$HR_hu_unpaired, ai.paired$HR_ai_paired, hu.paired$HR_hu_paired, var.equal=FALSE)
```

**Appendix L**

*Pearson's Correlation Between the Risk Assessment per Scenario (Q2.1.) and Team Diversity (Q4.2.)*

| Scenario (Q2.1.) | Team Diversity (Q4.2) | Pearson's r | P-value |
|---|---|---|---|
| **HR AI (S1V2)** | Age Diversity | .029 | .855 |
| **HR AI (S1V2)** | Gender Diversity | -.238 | .134 |
| **HR AI (S1V2)** | Race Diversity | .256 | .107 |
| **HR AI (S1V2)** | Disability Diversity | .286 | .070 |
| **HR AI (S1V2)** | Sexual Orientation Diversity | .339 | *.030 |

| | | | |
|---|---|---|---|
| **Bank AI (S2V2)** | Age Diversity | .277 | .083 |
| **Bank AI (S2V2)** | Gender Diversity | .017 | .919 |
| **Bank AI (S2V2)** | Race Diversity | -.072 | .660 |
| **Bank AI (S2V2)** | Disability Diversity | .143 | .380 |
| **Bank AI (S2V2)** | Sexual Orientation Diversity | -.039 | .811 |
| **Insurance AI (S3V2)** | Age Diversity | .343 | *.038 |
| **Insurance AI (S3V2)** | Gender Diversity | .173 | .306 |
| **Insurance AI (S3V2)** | Race Diversity | .415 | *.011 |
| **Insurance AI (S3V2)** | Disability Diversity | .194 | .250 |
| **Insurance AI (S3V2)** | Sexual Orientation Diversity | .119 | .484 |
| **Police AI (S4V2)** | Age Diversity | -.076 | .659 |
| **Police AI (S4V2)** | Gender Diversity | -.388 | *.019 |
| **Police AI (S4V2)** | Race Diversity | .111 | .521 |
| **Police AI (S4V2)** | Disability Diversity | -.108 | .530 |
| **Police AI (S4V2)** | Sexual Orientation Diversity | -.024 | .890 |
| **Prison AI (S5V2)** | Age Diversity | 4.041e -4 | .998 |
| **Prison AI (S5V2)** | Gender Diversity | .027 | .861 |
| **Prison AI (S5V2)** | Race Diversity | .042 | .788 |
| **Prison AI (S5V2)** | Disability Diversity | .299 | .051 |
| **Prison AI (S5V2)** | Sexual Orientation Diversity | -.122 | .436 |
| **Cancer AI (S6V2)** | Age Diversity | .112 | .481 |
| **Cancer AI (S6V2)** | Gender Diversity | .116 | .465 |
| **Cancer AI (S6V2)** | Race Diversity | .256 | .102 |
| **Cancer AI (S6V2)** | Disability Diversity | .276 | .077 |
| **Cancer AI (S6V2)** | Sexual Orientation Diversity | -.024 | .881 |
| **Covid AI (S7V2)** | Age Diversity | .029 | .854 |
| **Covid AI (S7V2)** | Gender Diversity | -.040 | .799 |
| **Covid AI (S7V2)** | Race Diversity | .208 | .181 |
| **Covid AI (S7V2)** | Disability Diversity | .026 | .870 |
| **Covid AI (S7V2)** | Sexual Orientation Diversity | -.031 | .842 |

*Note: *Significance level 0.05 is reached*