LUND UNIVERSITY
School of Economics and Management

Department of Informatics

Bias in the Context of Artificial Intelligence Systems:

Analyzing the risks and contributors from a data perspective

Master thesis 15 HEC, course INFM10 in Information Systems

Authors: Stefany del Carmen Firera Colmenares
Mana Vakil

Supervisor: Miranda Kajtazi

Grading Teachers: Osama Mansour
Saonee Sarker

Bias in the Context of Artificial Intelligence Systems: Analyzing the risks and contributors from a data perspective

AUTHORS: Stefany del Carmen Firera Colmenares and Mana Vakil

PUBLISHER: Department of Informatics, Lund School of Economics and Management, Lund University

PRESENTED: June, 2022

DOCUMENT TYPE: Master Thesis

FORMAL EXAMINER: Osama Mansour, PhD

NUMBER OF PAGES: 171

KEY WORDS: bias, data bias, artificial intelligence, risk of bias, algorithmic bias,

ABSTRACT (MAX. 200 WORDS):
As Artificial Intelligence (AI) is progressing to take over decision making in different industries, the threat that comes with the use of these systems is also increasing. One major threat is the risk of these systems acting biased, causing discrimination to parts of the population. To tackle the risk of AI systems acting biased it is important to understand how these biases originate in the first place. An analysis is made to understand where in the process the risk of bias takes place as well as highlighting the major contributor to bias in AI systems. Through a qualitative approach, practitioners currently working with AI and data were interviewed and presented with real-life examples of AI systems acting biased to help identify the reasons for the biased outcomes in these AI systems. The findings in this thesis indicate that data is a major contributor to bias in these systems, however, research has mostly been attributed to algorithms. Conclusively, this thesis found that there is a high risk of bias in the data collection, data preparation and model development stages in the AI systems.

Acknowledgements

Firstly, we would like to express extreme gratitude to our supervisor Miranda Kajtazi for all her invaluable support, feedback, and patience. We could not have asked for a better supervisor. The supervisions were essential to finding the right path in our research. With her input and enthusiasm, we were able to see and explore different points of views that made this research not just interesting but also exciting.

Secondly, we would like to thank all the participants who provided us with valuable insights, the rich interviews allowed us to see the subject through different perspectives.

Lastly, we would like to thank our families and friends for all their love and support during this period, especially during the long hours studying.

Stefany Firera & Mana Vakil

Content

Figures

Tables

1 Introduction

The following chapter will present an introduction to our research by giving the reader an overview of AI systems used to support and make decisions as well as exhibiting current problems that AI systems are facing regarding the risks of producing biased decisions. It will then continue by stating the purpose of this study and present the research question that is aimed to be answered as well as the limitations of the conducted research.

1.1 Background

During recent years there has been an increase in systems using Artificial Intelligence (AI) to improve everyday services, especially those of decision making (Kordzadeh & Ghasemaghaei, 2021; Lee, 2020; Ferrer, Nuenen, Such, Cote & Criado, 2021; Mikalef, Conboy, Lundström & Popovoc, 2022). For instance, systems using AI to influence decisions are used in many different fields such as human resources, finance, education as well as medicine (Baker & Hawn, 2021; Ferrer et al., 2021; Kordzadeh & Ghasemaghaei, 2021; Mehrabi, Morstatter, Saxena, Lerman & Galstyan, 2021). The output of these systems has the power to affect people's lives and if they are not being properly vetted they could lead to errors (Ferrer et al., 2021; Kordzadeh & Ghasemaghaei, 2021; Ntoutsi, Fafalios, Gadiraju, Iosifidis, Nejdl, Vidal, Ruggieri, Turini, Papadopoulos, Krasanakis, Kompatsiaris, Linder-Kurlanda, Wagner, Karimi, Fernandez, Alani, Berendt, Kruegel, Heinze, Broelemann, Kasneci, Tiropanis & Staab, 2020; Suresh & Guttag, 2021). Google AI (n.d) remarks that the rapid growth of AI is taking over sensitive industries, where a problem of bias could be extremely harmful and dangerous. It is important to start detecting bias as early as possible to mitigate the risks of problems affecting the systems (Leavy, 2018; Mehrabi et al., 2021)

AI systems have been found to show biases towards ethnicity, social groups, cultural backgrounds, age and gender (Mehrabi et al., 2021; Ntoutsi et al., 2020). The systems are not consciously biased by themselves, however, the decisions the systems make are based on the data that they are learning from as well as on the algorithms they are based on (Ferrer et al., 2021; Hellström, Dignum & Bensch, 2020). The former is when the system can produce biased assumptions by the use of inappropriate data or by wrongfully data preparation whereas the latter happens during the algorithm's unfairly development, evaluation, postprocessing or deployment (Ferrer et al., 2021; Hellström, Dignum & Bensch, 2020; Kordzadeh & Ghasemaghaei, 2021; Parikh, Teeple & Navathe, 2019; Suresh & Guttag, 2021). Since AI is being developed by humans, it is not only inheriting our own biases but also augmenting them (Ferrer et al., 2021; Mikalef et al., 2022; Ntoutsi et al., 2020). Ferrer et at., (2021) remark that when data sets used to train algorithms contain bias, these are likely to be reflected in the decisions of the systems. As evidence, inputs in some AI systems have been found to be biased (Mehrabi et al., 2021; Ntoutsi et al., 2020) which has led to wrongful outputs and discrimination (Ferrer et al., 2021).

As a result of AI increasing efficiency in the industries, AI market revenues worldwide are expected to rise by 15% in 2023 (Sujay, 2022). However, despite the advantages that AI systems provide, Mikalef et al., (2022) remark that it is necessary to evaluate AI systems from a critical point of view, as their faults are often shadowed by their benefits. For instance, the

systems have shown that they can be detrimental when making decisions such as hiring choices (Dastin, 2018), sentencing (Angwin, Larson and Kirchner, 2016), approving loans (Hassani, 2020) or determining grades (Baker & Hamn, 2021). As biases are being discovered by their creators or by the general public, companies and governments are starting to take action against those biases (Leavy, 2018; Mikalef et al., 2022). Examples can be seen in Google's responsible AI Best practices guidelines (Google AI, n.d.) and IBM initiatives for mitigating AI biases (Ferrer et al., 2021; Hobson & Dortch, 2019). Additionally, governments are creating policies on responsible AI to address AI challenges (European Commission, 2019; OECD, 2022). However, to fight this battle it is necessary to work preventatively, therefore understanding where the highest risk of bias is coming from is essential to avoid facing previous mistakes.

1.2 Problem

Algorithms learn from data and analyse patterns of the data. There has been a recent increase in talks about algorithms that are trained with biased data and their outputs being biased and discriminatory (Bolukbasi, Chang, Saligrama & Kalai 2016; Caliskan, Bryson & Narayanan, 2017; Ferrer et al., 2021; Hellström, Dignum & Bensch, 2020). In such circumstances, if existing and future systems continue to work on the basis of "poor" data - underrepresenting diversity, bias will only lead to many more problems. For instance, two of the world's tech giants, Apple and Amazon, have been accused of using AI systems that acted wrongfully on many occasions and that they still deal with loopholes.

To bring a better context, one recent example of AI systems acting wrongfully and getting accused of being gender biased was when it became known in 2019 that Apple's credit card issuer was not giving spouses the same amount of credit despite them having the same assets. It came to light that despite women in the relationship having the same assets as their husbands and in some cases even having higher credit scores, they were granted less credit when applying for a credit card (Nedlund, 2019). The report from the investigation stated that sometimes the data that is being used by creditors for developing and testing a model could result in unintended biased decisions being made (Nasiripour & Farrell, 2021).

Amazon also faced similar accusations with their AI hiring tool acting unfairly toward women, giving them inferior ratings in comparison to men (Dastin, 2018). The reason for this was that the data that was used to train the system was mostly from men's resumes, resulting in the system thinking that hiring men was the desirable choice. This shows how much impact the choice of data set will have on the outcome of the predictions and calculations that the AI systems make.

The problem is however not limited only to gender bias. There are examples of AI systems showing racial bias as well as other types of biases. One system that caught a lot of attention was COMPAS, which was used to determine the risk of a person re-offending. Dark-skinned people were more often predicted to re-offend (Corbett-Davies, Pierson, Feller & Goel, 2016).

Especially in the medical field, these biases can be dangerous, for example, with the help of AI systems, skin cancer can be detected, and the accuracy is recognized to be the same as when human experts diagnose the disease (Esteva, Kuprel, Novoa, Swetter, Blau & Thrun, 2017), which is a great move forward in technology. However, if the data sets do not

represent different skin characteristics such as different skin colors, skin textures & hairiness, the accuracy will only be high for the people that have their features represented in the data set (Buolamwini & Gebru, 2018).

In a discussion with Madkargav (2021), Bano mentions that whatever bias is embedded in the data sets will be integrated into the AI systems, meaning that if the data sets have sexism or racism embedded the outcomes of the predictions and decisions can be sexist or racist.The European Union's documentation treating AI policies shows that policymakers have not been able to capture the discriminations that can be caused by AI systems. AI systems are known to have acted wrongfully before and because of this, the need for a response is necessary (Balayn & Gürses, 2021).

Furthermore, AI is a technology that has and will continue to impact our businesses and society. Benbya, Pachidi and Jarvenpaa (2021), as well as Mikalef et al., (2022), argue that there is still a strong need for further research regarding ethical concerns on AI technology acting biased. They state that it is crucial to continue increasing the understanding of data practices and the risk of the systems leading to biased outputs. Caton and Haas (2020) and Kordzadeh and Ghasemaghaei (2021), confirm this by arguing that there is a fair share of literature addressing different approaches to mitigating bias and pushing for fairness but despite this, the subject is hard to tackle. Similarly, Zeng and Wu, (2021) continue by arguing that there is a need for diversification since the available literature is mainly focused on algorithms, whereas we focus on contributing by increasing knowledge regarding the data and the possible risks of places where bias can occur in AI systems.

AI is often portrayed as life-changing with many opportunities to make life better and easier. Despite this, researchers (e.g. Mikalef et al., 2022; Grewal, Guha, Satornino & Schweiger, 2021) state that special attention should be drawn to research on understanding where AI goes wrong, which is referred to as the dark side of AI. As for responsible governance of AI systems on a societal and regulatory level, there are various studies addressing the matter, however, there is a gap for empirical studies at an organizational level within the Information Systems (IS) discipline (Kordzadeh & Ghasemaghaei 2021; Mikalef et al., 2022). Furthermore, Mikalef et al., (2022) claim that the IS discipline should look more critically at the field of AI and its dark side. By turning to practitioners within the field of AI we are hoping to get a better understanding of data's impact on their systems. Additionally, we strive to identify where the highest risks of bias take place, thus, we are hoping to contribute to reducing the current knowledge gap and encourage safer AI systems.

Marabelli, Newell and Handunge (2021) point out that AI research often is funded by big tech corporations resulting in them having some influence on it, possibly making it hard to be objective and facing ethical problems. Data can be tough to be used responsibly when companies face situations where irresponsible use of data could lead to possible growth (Newell & Marabelli, 2015). One can therefore argue that research in the field with no conflicts of interest should be of high priority. Assuming these systems are not being used in a cautious way, the minorities and most vulnerable members of our society might end up getting hurt (Pilkington, 2019). With this being said, the need for further research regarding the data's effect on bias lays the basis for our research.

.

1.3 Purpose

Currently it is known that AI systems used to either support decisions or make decisions can lead to discrimination. This research focuses on the role of bias influencing the outcomes of AI systems decisions. The goal is to focus on this through a data bias perspective, leading to highlighting the importance of unbiased and high-quality data in AI systems as well as identifying where the risks of biases take place. With this focus at hand, this research identifies the key contributor of bias in AI systems, primarily by examining the literature, conducting, and interpreting the empirical findings and by using real-life examples of systems acting in a biased way.

In this vein, this research also attempts to identify where in the pipeline of the AI systems the biases are most likely to appear. Thus, demonstrating the importance of differentiating data bias from algorithmic bias.

1.4 Research Question

- What identifies as a key contributor to bias in AI systems, and where does the risk lay?

1.5 Delimitation

The word “AI systems” used in this thesis will refer to AI systems that are used to either support or make decisions. Furthermore, we recognize that biased decisions in AI systems can be harmful in several ways. Our study focuses on addressing bias on attributes such as gender, race, ethnicity, disability, religion, and age.

Moreover, despite there being different factors prompting bias in AI systems, this research paper will limit its scope to investigate this from a data perspective. Interviews have been conducted with participants that have worked directly with AI systems and have experience using data directly affecting these systems. Furthermore, we are choosing to focus on bias detection and the effects that biased data has on the systems rather than mitigating techniques to avoid these biases. We acknowledge that mitigation practices are very important topics to conduct research on however it is beyond our scope.

2 Literature Review

The following chapter presents findings from the literature that were considered relevant to this research. The chapter is divided into four subchapters focusing on Artificial Intelligence, The role of data in AI, Bias as part of the 'Dark side' of AI and lastly Bias in the development of AI systems. Additionally, eight real-life examples of AI systems acting biased are presented. Lastly a summary of this chapter is provided to the reader.

2.1 Artificial Intelligence

The term 'Artificial Intelligence' does not have a universal definition today, and it is defined in many different ways. Russell and Norvig (2016), point out that researchers in the field often argue that to define it, the first step is to define intelligence. The famous Turing Test, designed by Alan Turing in 1950 was an attempt to provide a definition of intelligence by providing a set of written questions. If the person providing the questions does not manage to determine if the responses are written by a computer or a human, the computer has reached intelligence (Haenlein & Kaplan, 2019; Russell & Norvig, 2016). There are, however, disagreements between researchers regarding if this qualifies as intelligence and often developers do not care whether or not it is real intelligence or just a simulation of it as long as their systems work as they intend them to. When talking about AI systems there are four combinations that are commonly mentioned, the system's ability to act humanly, the system's ability to think humanly, the system's ability to think rationally and lastly its ability to act rationally (Russell & Norvig, 2016).

AI goes further back than the Turing test. The main beginning of AI dates back to the 1940s (Haenlein & Kaplan, 2019), with the beginning of neural network research (McCulloch & Pitts, 1943). Despite it being considered an academic discipline, it was not until the Dartmouth Summer Research Project on Artificial Intelligence Conference in 1956, which was hosted by Marvin Minsky and John McCarthy, that AI started to be researched, with the broad goal of creating a human-like thinking system (Haenlein & Kaplan, 2019; Lee, 2020). The purpose of the conference was to bring researchers together to conduct research that would make it possible to embed human learning aspects to machines so they could mimic behavior together with the thinking process thus managing human problems (Lee, 2020).

Current AI has been defined as technologies that are able to simulate human ways to process, reasoning and learning (Lee, 2020)**.** Marvin Minsky, one of the most influential practitioners within AI and often referred to as "the father of Artificial Intelligence" (Haenlein & Kaplan, 2019; MIT MediaLab, 2016), defines AI as "…the science of making machines do things that would require intelligence if done by men" (Dennis, n.d.). A definition relevant to the field of IS was given by Mikalef and Gupta (2021), this definition gives emphasis to the AI goal of achieving predetermined organizational and societal objectives, by identifying, interpreting, making inferences, and learning from data. Due to the sociotechnical implications that can be referenced to IS, this definition will be used throughout this research.

There are a lot of different AI systems. One of them is recommendation systems which are often talked about as one of the most useful tools in today's digital world. Most of us encounter recommendation systems on a daily basis, often several times a day. These are reflected in our daily recommendations of things such as news articles, educational services and travel suggestions (Tsaku & Kosaraju, 2019). Furthermore, AI systems are also being widely used to influence decisions in areas such as criminal justice, child welfare, education, and immigration (Whittaker, Crawford, Dobbe, Fried, Kaziunas, Mathur, West, Richardson, Schultz & Schwartz, 2018). As previously mentioned, Amazon's hiring tool is an example of an AI system that was used for more efficient decision-making, however, it was discontinued due to the decisions being biased.

Today one can say that it is almost inevitable to not have daily interactions with AI systems. Haenlein and Kaplan (2019) argue that AI, if not already, very soon will have the same role in our lives as the internet has. Furthermore, the authors argue that it will be exciting to see how AI systems will be further integrated into our society and how they will coexist with humans.

2.1.1 Algorithms in AI

AI systems have different algorithms in their system that work by finding and learning different patterns from the historical data that they use as input (Suresh & Guttag, 2019). An algorithm is described as a process containing several sequences of instructions, one needs to complete for finalizing a specific task (Garcia, 2016), however, some steps may be performed only if the conditions require it (Mueller & Massaron, 2021; Skiena, 2020). Algorithms have been used for many years, however, before the emergence of technology algorithms were performed manually taking a lot longer. Today, computers, together with the help of different algorithms can find solutions to many problems within a reasonable time and in the easiest way possible. An algorithm can contain a lot of different operations such as data storing, exploring the data, and fixing the structure of the data. They then generalize their findings and apply them to new unseen data to produce outputs (Suresh & Guttag, 2019). Algorithms are widely used and can be found in most fields such as science, finance and medicine to name a few (Mueller & Massaron, 2021).

Algorithms are an essential part of today's technology (Garcia, 2016), and could easily be described as the artifacts that are used in computer programs to help with tasks such as data processing (Hill, 2016). Algorithms in AI can be used for many different tasks such as deciding on who to tag in photos or the reply you get when talking to virtual assistants (Garcia, 2016).

2.2 The Role of Data in AI

2.2.1 The data journey

The role of data in AI is crucial for the systems to work, as raw data provides the necessary input that can help AI systems make decisions. Databases are constantly increasing in size due to automation of transactions, immediate access to information via the internet and the expansion of storage, thus making room for even more data production (Fernandez, Garcia, Galar, Prati, Krawczyk, Herrera, 2018; Frawley, Shapiro & Matheus, 1992;). Each transaction

produces data and the produced data is registered and stored (Frawley, Shapiro & Matheus, 1992; García, Luengo & Herrera, 2015).

Data can be collected in different ways, for instance, data can be taken in the form of social data which can include data from social media, collaboration, or blog platforms. Online websites provide access to a great amount of data with a high level of details on users' everyday life (Olteanu Castillo, Diaz & Kiciman, 2019). In order to analyze the data collected, proper techniques should be performed to prepare the data**.** Preparing the data can help improve its quality, if the data quality is low, then the whole system will be affected (García, Luengo & Herrera, 2015).

To prepare the data, a great deal of effort must be put into the data preparation or pre-processing stage. This stage handles data collected that presents errors, inconsistencies, lack of specific trends and is incomplete (Fernandez et al., 2018). Data preprocessing are proven techniques that can solve issues involving the data by converting raw data into a comprehensible format (Lee, 2020). Data preprocessing basic techniques include cleaning, integration, normalization, transformation and reduction of data to be able to produce a data set that meets the requirements for high data quality (García, Luengo & Herrera, 2015; Lee, 2020). To elaborate on the techniques one can start by discussing data cleaning, which is an approach that tackles incomplete data, for instance, dealing with insufficient attributes or missing values (Lee, 2020). Data cleaning can clear errors and noise in the data as well as clear data that shows discrepancies (Han, Kamber and Pei, 2012; García, Luengo & Herrera, 2015). Data transformation is the next step in the preprocessing techniques, in this step, the data that has been collected in a different format is converted and unified to match the specific format requirements (García, Luengo & Herrera, 2015; Lee, 2020).

Following data integration, the technique merges data that has been taken from a variety of sources into a legible collection of data. Data can be delivered in different types and formats; therefore, it is necessary to ensure consistent collection of data that can manage redundancy and inconsistencies across data sets. This can be done with the help of data integration (García, Luengo & Herrera, 2015; Lee, 2020). A positive aspect of data integration is that it enables a more unified format, facilitating access to more efficient and precise data (Lee, 2020). Furthermore, data normalization involves arranging data within a database so that redundancies can be reduced, and data integrity can be enhanced (Lee, 2020). Finally, data reduction can create a smaller sample of the data set maintaining its structure and integrity to produce comparable analytical results (Han, Kamber and Pei, 2012; García, Luengo & Herrera, 2015).

Problems presented in the raw data can occur due to human or computer errors, transaction issues or defective instruments (Lee, 2020). It is necessary to apply the proper techniques to be able to create improved data sets that can be used in the analysis processes. García, Luengo and Herrera (2015) point out that the data preparation needs to be handled with proper care, otherwise algorithms running on this data will report incorrect results. Similarly, Han, Kamber and Pei (2012) note that improved data quality can be achieved through preprocessing techniques**.** The quality of how these steps are performed is what determines the outcome of the decision and its quality (Ge, 2019).

Once the data is prepared and ready, the next step is to analyze it. Often data sets are being divided, allocating a portion to the training data that is being used in model development, and test data that can be used in the evaluation of the model. Additionally, a portion of the data being used for training can be utilized as validation data (Suresh & Guttag, 2021). To teach the model a training set of data is necessary, consecutively in order to determine the accuracy

of the model, the test set is being used. Furthermore, the purpose of a validation set is to locate and optimize the most suitable model for the system (Russell & Norvig, 2021). Different methods can be applied to analyze the data, such as machine learning or data mining. These approaches can explore and find patterns in data sets utilizing intelligent methods to be able to produce an output that can be used in the system (Han, Kamber and Pei, 2012 García, Luengo & Herrera, 2015; Lee, 2020; Russell & Norvig, 2021).

2.2.2 Data quality

Data quality is an important part of the operating process for decision making as data is being used as the base for decisions, thus having data of high quality as inputs should lead to qualitative decision outcomes (Batini, Cappiello, Francalanci & Maurino, 2009; Han, Kamber & Pei, 2012). However, most of the time, the data collected lacks quality, some reason for this could be that the raw data is gathered from multiple sources and can be presented in different formats thus leading to errors, discrepancies or incomplete data (García, Luengo & Herrera, 2015). These issues should be able to be revised by the elements in the data preprocessing thus enhancing quality data for data analysis purposes (García, Luengo & Herrera, 2015; Han, Kamber & Pei, 2012). However, in order to ensure data quality for values, dimensions should be considered. Batini et al., (2009) remark that there are no general guidelines nor definitions for dimensions to determine high data quality, however, there is a basic set of main dimensions which are accuracy, completeness, consistency and timeliness. Additionally, believability and interpretability can be included due to the value they currently represent (Han, Kamber & Pei, 2012; Lee, 2020).

Accuracy can be defined as the level of similarity between the data value presented and its real value. Accuracy in a database can be measured by the number of faulty values it contains in order to check whether the data is correct and reliable (Batini et al., 2009; Fox, Levitin & Redman, 1994; Lee, 2020). Similarly, consistency indicates if related values in the same database are constant (Fox, Levitin & Redman, 1994; Lee, 2020). Following, completeness refers to the extent to which all attributes that are supposed to have values in the data collection have those values, meaning that there should not be missing values on specific attributes (Fox, Levitin & Redman, 1994; Batini et al., 2009). Details can be subject to change depending on the time (Fox, Levitin & Redman, 1994), timeliness refers to having data updated over time, as data is sometimes required during a particular period (Batini et al., 2009; Lee, 2020; Han, Kamber & Pei, 2012).

Batini et al., (2009) explain that with data complexity increasing, there is also a need to update and add new techniques to data quality. Thus continuing with two previously mentioned dimensions of data quality which are believability and interpretability. For instance, believability considers the degree to which people trust the data, tackling the question of data credibility (Han, Kamber & Pei, 2012; Lee, 2020) and interpretability indicates the ease with which data can be understood (Han, Kamber & Pei, 2012).

Olteanu et al., (2019) explain that attaining quality is desirable, however when using great amounts of data, full data quality is not always achievable. As there is a growth in the systems using data, there is also a need to supply more quality data to these systems (Batini et al., 2009). Nevertheless, Han, Kamber and Pei (2012) remark that the quality of the data relies on its purpose, for instance, employees with different roles might regard the quality of the data differently.

2.3 Bias as part of the 'Dark side' of AI

AI's many benefits can sometimes overshadow the risk that using this technology might expose communities to. A few examples of the negative effects of AI are data privacy concerns, lack of transparency, as well as biased outcomes. These negative effects of AI have been dubbed the 'dark side' by researchers (Cheng, Lin, Shen, Zarifis & Mou, 2022; Grewal et al., 2021; Mikalef et al., 2022; Zeng & Wu, 2021). To a further extent on the systems acting biased, AI systems have been demonstrated to contain bias, thus producing harmful outcomes (Grewal et al., 2021; Mikalef et al., 2022). As a result, social inequalities are likely to increase due to these biased outcomes.

The Cambridge Dictionary (2022), 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 judgment"**.** Additionally, bias could be referred to as the conclusion made based on prejudices, rather than impartiality (Hellström, Dignum & Bensch, 2020). In the context of data, bias means that the data used is distorted in a systematic way, compromising its representativeness (Olteanu et al., 2019). Additionally, Mitchell, Potash, Barocas, D'Amour, and Lum (2021) define bias in AI systems as obtaining an outcome out of the model that can be considered unjust towards disadvantaged groups that share specific attributes such as race, gender and others. However, Ferrer et al., (2021) observe that bias might not always be bad, sometimes it is necessary in order to find differences between events and patterns in data, as they can be referred to as variations from the standard. Nevertheless, Suresh and Gutta (2021) point out that there are often possibilities that biases will cause flaws in parts of the AI systems**,** emphasizing that data properties can cause unintended consequences favoring other members of society.

Furthermore, the term bias in AI systems has been widely discussed making reference to unfairness (Mehrabi et al.,2021). Similar to the term bias**,** unfairness in the context of AI systems is described as indulging in unequal treatment or exhibiting prejudices towards either individuals or groups based on specific attributes (Mehrabi et al., 2021). Decisions taken from an unfair algorithm can be misleading thus favoring some groups more than others, in some cases even ignoring specific groups thus increasing social inequities **(**Holstein, Wortman Vaughan, Daumé, Dudik & Wallach, 2019).

2.3.1 Bias and Discrimination

The concept of bias is broad as a consequence of the term being covered in different fields such as policy, law as well as ethics (Ferrer et al., 2021; Hellström, Dignum & Bensch, 2020; Holstein et al., 2019). As a result of the broad definitions of bias, there is a special emphasis on social biases in data affecting AI systems. This refers to the lack of people's characteristics representing those characteristics of a broader population (Ferrer et al., 2021; Olteanu et al., 2019). The demographic attributes used in data sets can be those of race, gender, age, income, sexual orientation and education. The mentioned biases, go hand in hand with discrimination practices as bias against people can lead to disadvantages affecting people's lives. Discrimination refers to the unjust considerations of people based on the mentioned demographic attributes (Ferrer et al., 2021). Discrimination in systems can be referred to as digital discrimination, amplifying current social inequalities through systems for example by affecting hiring or criminal convictions**.**

Barocas, Hardt and Narayanan (2019) discuss discrimination as the consequence of encountering bias in the AI systems by either data or model errors. Suresh and Gutta (2019) argue that the use of these attributes plays a different role depending on what the system is being used for, for example, attributes relating to gender or ethnicity should not be relevant to hiring decisions, however, in a medical context these attributes do play a significant role in determining patient illness.

2.4 Bias in the Development of AI Systems

AI is often the reflection of high amounts of observations of a given context. These observations are often taken from the real world, thus the results of the systems can be as close to reality as possible. However, the real world often contains biases, which are then replicated by the systems (Hellström, Dignum & Bensch, 2020; Suresh & Guttag, 2019). Mikalef et al., (2022) mention that careful attention should be paid to the data as well as the training of the algorithms to ensure there is no bias in the process. Likewise, Ferrer et al., (2021) observe that in order to assess if a system is biased, there is a need to perform an analysis of the whole process, from understanding the data and the algorithms underlying assumptions and models to the context of the algorithm's decisions and developers' prejudices. However, Ferrer et al., (2021) and Holstein et al., (2019) note that one does not always have access to all this information. For instance, data or algorithms used for the systems can be protected, due to the amount of confidential information they contain.

Biases in the data can be presented in the systems by the data generation process that involves data collection as well as data preparation (Baker & Hawn, 2021, Suresh & Guttag, 2019; Suresh & Guttag, 2021). Bias in the data collection can be the result of the selection or handling of data sources that could contain bias. Olteanu et al., (2019) further explain that the bias in the data preparation refers to data sets that can be affected by the way their content is being processed after being collected.

Biases in algorithms can originate in different stages such as model development, model evaluation, model postprocessing and model deployment (Baker & Hawn, 2021; Suresh & Guttag, 2021). Model development involves the definition and creation of a particular model with an objective function; model evaluation requires the evaluation of a developed model with a test data set; model postprocessing manages the steps after training in order to achieve the desired output and lastly, model deployment involves launching a model for a specific use.

Balayn, Lofi and Houben (2021) note that machine learning processes are often used for the current development of AI systems to support decision making, which is trained and evaluated using data sets. Following Suresh and Guttag's (2021) approach to identifying bias, the various stages of the process are categorized as data collection, data preparation, model development, model postprocessing, model evaluation and model deployment. In the interest of the study, the place where bias can be found are divided into two different categories, which are data bias and algorithmic bias. The stages concerning data collection and preparation go under data bias and the stages concerning model development, postprocessing, evaluation and deployment go under the umbrella of algorithmic bias.

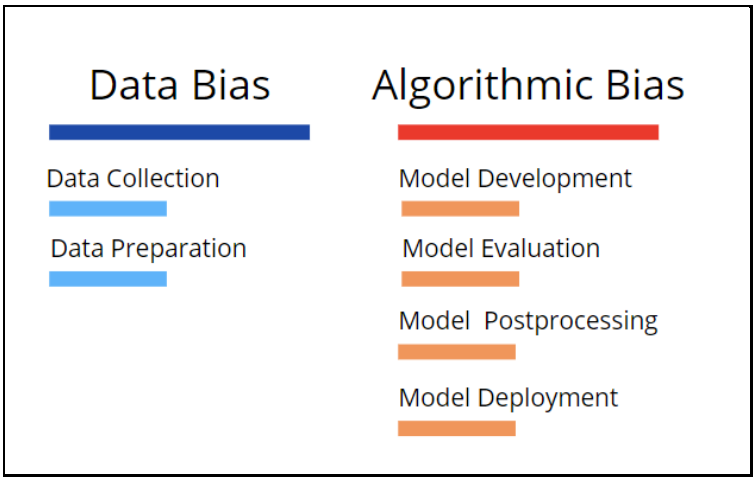


Figure 1: Stages in data and algorithmic bias

It is important to highlight that biases in AI systems can be caused by the data used for training, the algorithms, as well as the professionals working with the systems (Ferrer et al., 2021; Hellström, Dignum & Bensch, 2020; Olteanu et al., 2019). Below sources of bias in the AI systems can be found.

2.4.1 Data Bias Affecting AI Systems

It is not unusual for data to reflect the biases that are embedded in our society today (Barocas & Selbst, 2016). Baker and Hawn (2021), reflect on how a large fraction of biases discovered in algorithms are a result of factors outside the model's actual training process. These factors are directly involved with data and can range from the use of historical data to data representation and measurement characteristics. Holstein et al., (2019) highlight people's concern about bias affecting data directly rather than the algorithms.

Bias in data can be introduced to the systems when it is being used for training. For example, Olteanu et al., (2019) explain that data sets taken from social media sources are influenced by people's demographics on specific platforms, thus all relevant data might not be available. Baker and Hawn (2021) mention that the right data should be collected in order to create a more effective system. However, Suresh and Guttag (2021) observe that in practice employees are not always able to collect their own data, leading them to acquire external data sets. Data preparation uses preprocessing techniques in order to prepare the data to be used in the model. Actors contributing to bias in data preparation can be those generated by data preprocessing techniques such as data cleaning, normalization or transformation, as well as labelling (Suresh & Guttag, 2021). Different sources of bias in data sets are explained below:

2.4.1.1 Historical Bias

Even though the data sets that are being used are often picked with high caution the data that is included is often a simulation of the world as it is today, therefore the biases and problems that are embedded in today's society are reflected in the data. Historical biases are being introduced to AI systems due to unwanted attributes that have been collected over time (Ferrer et al., 2021; Hellström, Dignum & Bensch, 2020). Suresh and Guttag (2019), give an example of image searching, when Fortune Magazine published their 'Fortune 500' in 2018 only 5% of the CEOs were women (Zarya, 2018). Google's image search displayed fewer women than men when one googled 'CEO'. This is a reflection of the historical data, showing that

historically men are more likely to hold CEO positions, therefore the search engine displayed more pictures of men.

2.4.1.2 Representation bias

Representation bias occurs when the data set is not inclusive enough, meaning that certain groups are underrepresented or overrepresented in the data set. There are different reasons contributing to representation bias. One of them is that their sampling model is not reachable enough, for instance, a lot of today's data is gathered from smartphones and different smartphone applications. This is an effective and easy way of collecting data, however, elderly population that do not know how to use these technologies will not be included, and neither will the population of the countries that do not have the same access to smartphones and the internet. Representation bias in this context is usually referred to as sampling bias in statistics (Suresh & Guttag, 2021). Another reason could be that the target group has changed. Data that was representative of one city does not have to work as well for another city (Suresh & Guttag, 2019). Baker and Hawn (2021) emphasize that data collected should involve all samples targeted by the model, otherwise one should not expect the models to work on everyone.

2.4.1.3 Measurement bias

Depending on how one measures and utilizes different features one can encounter measurement bias. The data that is measured and available is often proxies for some other labels. To make it easier to understand the authors, Suresh and Guttag (2019), give the example of the arrest rate which often is the proxy for crime rate. This occurs when one does not manage to measure what is intended to be measured correctly. Some reasons for measurement bias occurring can be that the granularity of data is different. An example of this is the COMPAS software that used the prior arrest status of the defendant's family and friends to determine the likelihood of them re-offending. In the case of COMPAS information about the defendant's family and friends' prior arrests were used as proxies for the measure of their risk to commit a new crime. It is not unusual for minority communities to have a high police presence, leading to a higher number of arrests. Drawing the conclusion that the residents of minority communities have a higher number of dangerous people due to having a higher number of arrests can however be misleading due to factors such as the police presence. Another reason could be that the quality of the data is different for different groups (Suresh & Guttag, 2019). Baker and Hawn (2021) provide another example mentioning that training labels can be biased, making reference to labeling dark-skinned students as being more likely to provoke violence at school even though they have engaged in the same behavior as white students. If this type of mislabeling occurs the algorithm will not be able to differentiate between biased or fair labels.

2.4.2 Algorithmic Bias Affecting AI Systems

Hellström, Dignum and Bensch (2020) remark that bias could originate from the model processes, where the objective is to use it to learn from a data set and create outputs similar to the ones expected. Some bias might be added to the systems in order to compensate for bias found in the data (Ferrer et al., 2021). Hellström, Dignum and Bensch (2020) consider that some bias might be necessary for the systems to work on specific parameters such as probability. For instance, even though all requirements to get a loan approved might be met, the loan might still be rejected due to the probability of it being approved is below the threshold.

Furthermore, some biases might be necessary depending on the context of the systems (Ferrer et al., 2021; Hellström, Dignum & Bensch, 2020). For instance, an example of bias in usage could be if a job requires young people, a biased algorithm could be used to exclude the elderly, which does not necessarily mean discrimination against older people as there is a justification.

2.4.2.1 Aggregation Bias

When one uses a "one-size-fits-all" model despite different groups having different conditional distributions there is a risk of aggregation bias. For instance, variables can have different meanings depending on the person (Suresh & Guttag, 2019). A common example of this is clinical aid tools that are used for patients with diabetes. There are many disparities in complications associated with diabetes depending on the ethnicity of the patient (Spanakis & Golden, 2013). Therefore, a "one-size-fits-all" model could result in aggregation bias.

2.4.2.2 Learning Bias

Different modeling choices tend to increase performance imbalance across data sets, causing learning bias. For instance, when defining objective functions where accuracy can be measured, prioritizing an objective could cause harm to another. Suresh and Guttag (2021), argue that prioritizing privacy in the training stage can lower the chances of underrepresenting data appearing in the model. Likewise, when data is applied to a compact model with underrepresented attributes, it can amplify inconsistencies in model performance.

2.4.2.3 Evaluation Bias

Baker and Hawn (2021) point out that evaluation bias is the result of using data sets for testing that does not reflect the target population for which the model was intended. Suresh & Guttag, (2019) explain that when optimizing models, one often uses its training data, however, for quality measurements benchmarks are used. If the model evaluation is done by using disproportionate benchmarks or benchmarks that are not suitable for its intended use, there is a risk of evaluation bias. An example could be using a benchmark that is not suited for the target group leading to the development of a model that is only suitable for a certain group. Furthermore, Suresh & Guttag, (2019), argue that evaluation bias occurs when there is a need for comparing different models. Additionally, it is pointed out that representation bias should be discovered in the benchmarking of the model otherwise they risk making the evaluation of the model biased as well.

2.4.2.4 Deployment Bias

This form of bias arises when the systems are deployed with a different purpose than they were intentionally designed for (Baker & Hawn; 2021; Suresh & Guttag, 2021). Additionally, this kind of bias often occurs when humans are involved in the decision-making process and use systems to confirm their own judgments. Baker and Hawn (2021) provide the example of a model being used for identifying student engagement and being deployed to allocate grades instead. Likewise, Suresh and Guttag (2021), highlight the example of the COMPAS algorithm, where the model was designed to predict the likelihood of someone carrying out a crime in the future, instead of being used for determining how long a sentence should be. Both examples can cause consequences as the systems were not designed to deal with different objectives.

2.4.3 Human Bias Affecting AI Systems

The development of AI systems for decision making is heavily influenced by humans because humans are responsible for bias; they have a role to play in introducing potentially biased systems to the market (Hellström, Dignum & Bensch, 2020). Baker and Hawn (2021), point out that humans can not only influence the model but also the data. When doing data preparation and training the models, labels might be assigned from a personal point of view rather than an objective one.

Additionally, it is important to notice that the limited diversity in the field of developers could be a reason for bias. In a discussion with Madkargav (2021), Bano mentions that 80% of the employees at the big machine learning companies possessing technical roles are males. Bano brings up a report conducted by UNESCO (2019) that shows that only 12% of the researchers in the field of AI are women and only 6% of the software developers that work with AI are women. When writing the algorithms there is a risk of developers transferring their biases into the algorithm. Equally, when processing data, data scientists can ingrain their beliefs into the systems (Ferrer et al., 2021).

Holstein et al., (2019) note that it is necessary for practitioners and researchers to support each other in the movement against bias in AI systems. This can be done by supporting people in the industry to collect and curate data as well as developing fairer algorithmic models. Garcia (2016), argues that all organizations and institutions that use algorithms should understand that diverse teams play a role in better detecting and anticipating problems the algorithms may present. Hankerson, Marshall, Booker, El Mimouni, Walker, and Rode (2016), argue that the absence of diversity during the different stages of the process can result in problematic systems outcomes, giving the example of the lack of people of color in the technology sector could lead to racial bias in systems.

2.5 Examples of AI acting biased

Mikalef et al., (2022) observe that the mentioned focus on more ethical, fairer and transparent AI systems comes as a consequence of AI's past biased outcomes. Mikalef et al., (2022) point out that it is important to comprehend where the negative effects of bias originated in the systems. The following section presents some real-life examples of AI acting biased in different kinds of systems. The cases are shown in a table to easily get an overview of AI systems' discriminative outcomes and the reason for it. The table is followed by a short explanation of all the cases.

Table 1: Overview of AI Systems acting biased

Case	Example	Type of bias	Reason for bias	Source
1	New Zealand passport robot rejected an Asian man's eyes because subject's eyes are closed	Racial bias	Data	(Griffiths, 2016), (Hellstrom, Dignum & Bensch, 2020)
2	Women seeing fewer job ads for STEM than men	Gender bias	Algorithm	(Dunphy, 2018), (Lambrecht & Tucker, 2018)
3	Racial Bias found in a major health care risk algorithm	Racial bias	Data	(Obermeyer, Powers, Vogeli & Mullainathan, 2019).
4	Amazon hiring tool acting in favor of men	Gender bias	Data	(BBC, 2018), (Hellstrom, Dignum & Bensch, 2020), (Dastin, 2018)
5	COMPAS - predicts the risk of someone re-offending	Racial bias	Data & Algorithm	(Angwin, Larson and Kirchner, 2016), (Corbett-Davies et al., 2016), (Saxena, Huang, DeFilippis, Radanovic, Parkes & Liu, 2018) (Suresh & Guttag, 2019)
6	Apple Credit Card acting gender biased	Gender Bias	Data	(Nasiripour & Farrell, 2021), (Nedlund, 2019), (Telford, 2019),
7	Family screening tool - abusive families	Socio-economic bias	Data	(McKenna, 2019)
8	Google's photo app labeled dark skinned couple wrongfully	Racial bias	Data	(Balayn, Lofi and Houben, 2021), (BBC, 2015)

Case 1: A man of Asian descent wanted to apply for a new New Zealand passport online. The photo he uploaded was rejected and he was presented with an error message that said his eyes were closed and the photo did not fulfil the requirements of the system. His eyes were open and after three attempts with three different photos, he had to reach out to the passport office. The office blamed it on bad lighting and shadows in the eyes (Griffiths, 2016; Hellstrom, Dignum & Bensch, 2020). This is a typical example of representation bias that occurred due to an underrepresentation of people of Asian descent in the data set used to train the model (Hellström, Dignum & Bensch, 2020).

Case 2: In 2018 it became known that ads related to STEM careers were less likely to appear for women than men. The ads were run on many big different social media platforms such as Facebook and Instagram. However, women were less exposed to the ads, and they were appearing more frequently to men (Dunphy, 2018; Lambrecht & Tucker, 2018). A possible reason for this could be that the algorithm was optimized to be cost-effective, resulting in it presenting the ads more frequently to men since displaying ads to men costs less (Lambrecht & Tucker, 2018).

Case 3: A well-used algorithm demonstrated racial bias. This was due to the algorithm making its calculations based on data that was not representative enough. The algorithm was used to predict what patients would find "high-risk care management" useful. To do this the system used healthcare data from previous patients that included their previous spending on healthcare. The risk scores that were received by dark-skinned people were often lower. This can be due to unequal access to healthcare resulting in dark-skinned people not seeking help as frequently which leads them to not be as represented in the data (Obermeyer, Powers, Vogeli & Mullainathan, 2019).

Case 4*:* The AI hiring tool that Amazon used demonstrated gender bias when recommending candidates for hire. The goal of the system was for it to rate candidates with a number between 1 and 5. It became known that the data sets that were used to train the model mainly consisted of male resumes resulting in the system teaching itself that recommending a male would be the desirable thing to do (Dastin, 2018; BBC, 2018).

Case 5: An algorithm called COMPAS was used in the United States to determine whether or not a defendant that is awaiting trial will re-offend. The decision is based on more than 100 different factors such as age and criminal history. When analyzing the results it demonstrated that dark-skinned people were more often classified as high risk among the defendants that did not commit new crimes (Corbett-Davies et al., 2016). A possible reason for this could be that minority communities have a high police presence, leading to a higher number of arrests. Drawing the conclusion that the residents of minority communities have a higher number of dangerous people due to having a higher number of arrests is misleading due to factors such as the police presence (Marabelli, Newell & Handunge, 2021; Suresh & Guttag, 2019). Other possible reasons could be the granularity of the data being different (Suresh & Guttag, 2019) as well as the fairness metrics being used that would favor one subgroup over another (Saxena, Huang, DeFilippis, Radanovic, Parkes & Liu, 2018)

Case 6: Apple's credit card issuer was accused of acting gender-biased when it became known that their system was not giving spouses the same amount of credit despite them having the same assets. Even though the women in the relationship had the same assets as their husbands

and in some cases even had higher credit scores were granted less credit when applying for a credit card (Nedlund, 2019; Telford, 2019). The investigation stated that sometimes the data that is being used by creditors for developing and testing a model could result in unintended biased decisions being made (Nasiripour & Farrell, 2021).

Case 7: This is a tool that has been used to help humans determine if a child should be relocated because of abusive circumstances in their homes. The model is based on public data sets leading it to reflect the widespread societal biases that we have embedded into our society today. This was favorable to middle and upper-class families since they could easily hide the abuse by using private health care. Their abusive conditions were not as widely represented in the data set. However, to mitigate the risks the model was designed transparently and was only used to help employees and did not make any decisions alone (McKenna, 2019).

Case 8: Google's photo app made a big mistake when using AI to automate photo labeling and ended up labeling a dark-skinned couple as gorillas. The company apologized for this but a lot of questions were triggered regarding what kind of data was used that led the AI to label the couple this way (BBC, 2015; Balayn, Lofi and Houben, 2021).

Baker and Hawn (2021) believe that often biases go unnoticed, however, once they are discovered they go from unknown bias to known bias. Balayn, Lofi and Houben (2021) stress that bias often goes unnoticed unless a system already deployed, acts unfairly towards a section of the population. Only when research is invested in understanding the bias, is the immensity of the problem discovered. Once a bias is discovered, measures can be taken to control it and avoid future biases in the systems thus taking one further step toward inclusive AI systems.

2.6 Summary of the literature review

In this section, a short summary of the above-mentioned literature will be provided.

As mentioned, the beginnings of AI date back to the 1940s and have grown tremendously, being expected to have the same role in our lives as the internet has in the not distant future (Haenlein & Kaplan, 2019). Thanks to the successful development of technologies the amount of collected data has increased dramatically (Fernandez et al., 2018; Frawley, Shapiro & Matheus, 1992). AI systems run with algorithms and data and they find and learn patterns from data (Suresh & Guttag, 2021), meaning that data is essential for many of its operations (Fernandez et al., 2018; Frawley, Shapiro & Matheus, 1992).

The data journey starts with data collection and making improvements by the data preparation in order to achieve high-quality data. Data not living up to a certain quality can affect the entire process (García, Luengo & Herrera, 2015). To increase the quality of data, techniques such as preprocessing can be done (García, Luengo & Herrera, 2015; Lee, 2020). The usage of bad quality data will lead to poor results in the analysis that are performed, which will in turn lead to failed AI systems or bad and low-quality decision-making (Batini et al., 2009; Han, Kamber & Pei, 2012). The literature offers dimensions for how to assure high data quality (Batini et al, 2009).

Furthermore, bias as part of the dark side of AI is explained, emphasizing the societal prejudices that might be embedded in the data used for the systems affecting the system's ability to

make fair decisions (Ferrer et al., 2021; Olteanu et al., 2019). The unfair decisions that are at risk of being made through AI systems will amplify the social inequities that we already see in our society (Ferrer et al., 2021). This means that real-world bias will be replicated by the systems since it can be integrated in the data (Hellström, Dignum & Bensch, 2020; Suresh & Guttag, 2019). Bias in data can be presented in two stages, during the data collection, the data preparation or both (Suresh & Guttag, 2019; Olteanu et al., 2019). Bias in the models can be introduced in four stages, model development, postprocessing, model evaluation and lastly model deployment (Suresh & Guttag, 2021). Additionally, when developing AI technologies, there is always a risk that the practitioners transfer their own biases to be reflected in the systems. Having a diverse team could reduce this risk (Hellström, Dignum & Bensch, 2020)

Independent of the cause of the bias, the biased decisions made by AI systems might lead to digital discrimination against people with different backgrounds. It is believed that because algorithms are learning from data if there is not a proper representation of people's characteristics in the data, the algorithms can provide biased algorithmic decisions (Ferrer et al., 2021). Barocas and Selbst (2016) state that if the algorithm is fed with bad data, one should not expect good quality outputs. To better understand the possible risks of an algorithm acting biased Suresh and Guttag (2019; 2021), argue that one should try to understand the process of the data generation and the model process that led to the biased output being produced. Knowing the systems can be helpful in detecting where the bias occurs. One can conclude that to tackle the risk of bias, it is important to understand the entire process, from the data collection to the model deployment.

Table 2: Summary of literature review

Concept	Aspect	References
Foundational concepts		
Artificial Intelligence	- Artificial Intelligence history - Algorithms	(Haenlein & Kaplan, 2019) (Russell & Norvig, 2016) (McCulloch & Pitts, 1943) (Lee, 2020) (MIT MediaLab, 2016) (Dennis, n.d.) (Tsaku & Kosaraju, 2019) (Siles, Segura-Castillo, Solís & Sancho 2020) (Patibandla, Tummalapalli, Lingamaneni, Prasanna & Kumar, 2021) (Whittaker et al., 2018)
The role of Data in Artificial Intelligence	- The data journey - Data collection - Data preprocessing - Data Analysis - Data quality	(Frawley, Piatetsky-Shapiro & Matheus, 1992) (García, Luengo & Herrera, 2015), (Borana, 2016) (Lee, 2020) (Fernandez et al., 2018) (Han, Kamber & Pei, 2012) (Frye & Heinrich, 2020) (Brownlee, 2020) (Brink, Richards & Fetherolf, 2016) (Fox, Levitin & Redman, 1994) (Batini, Cappiello, Francalanci & Maurino, 2009)
Bias as a part of the 'Dark side' of AI	- Dark side of AI - Bias - Discrimination - Demographic attributes	(Zeng & Wu, 2021)(Grewal et al., 2021) (Cheng, Lin, Shen, Zarifis & Mou, 2022) (Mikalef et al., 2022) (Cambridge Dictionary, 2022) (Hellström, Dignum & Bensch, 2020) (Olteanu et al., 2019) (Mitchell et at., 2021) (Ferrer et al., 2021) (Suresh and Gutta, 2021) (Mehrabi et al.,2021) (Holstein et

		al., 2019) (Barocas et al., 2019) (Suresh & Gutta, 2019)
Bias in the process for AI systems		
Bias in the development of AI systems	- Data collection - Data preparation - Model development - Model postprocessing - Model evaluation - Model deployment	(Suresh & Guttag, 2019) (Hellström, Dignum & Bensch, 2020) (Suresh and Guttag, 2021) (Mikalef et al.,2022), (Ferrer et al., 2021) (Holstein et al., 2019) (Baker & Hawn, 2021) (Olteanu et al., 2019) (Balayn, Lofi and Houben (2021)
Data bias affecting AI systems	- Historical bias - Representation bias - Measurement bias	(Barocas & Selbst, 2016). (Baker and Hawn, 2021) (Holstein et al., 2019) (Olteanu et al., 2019) (Suresh & Gutta, 2019) (Suresh and Gutta, 2021) (Hellström, Dignum & Bensch, 2020) (Zarya, 2018).
Algorithmic bias affecting AI systems	- Aggregation bias - Learning bias - Evaluation bias - Deployment bias	(Hellström, Dignum & Bensch, 2020) (Ferrer et al., 2021) (Suresh and Gutta, 2021) (Suresh & Gutta, 2019) (Spanakis & Golden, 2013) (Holstein et al., 2019)
Human bias affecting AI systems	- Human influence - Diversity	(Hellström, Dignum & Bensch, 2020) (Baker and Hawn, 2021) (Madkargav, 2021) (UNESCO, 2019) (Hankerson et al.,, 2016)(Garcia , 2016) (Holstein et al., 2019) (Ferrer et al., 2021)

3 Research Methodology

The following chapter will describe what kind of research design has been carried out for this study. Starting with the research philosophy that has been chosen, followed by the research approach, the methods that will be used for data collection and analysis, in addition to these, ethical considerations and scientific quality will be discussed.

3.1 Research philosophy

For this research we aimed to understand practitioners' points of view regarding the origins of bias in different kinds of systems using AI. Due to the social origins of biases, this study was conducted with the philosophical principles of interpretivism. Since interpretivism focuses on social constructs as a way of understanding reality (Myers, 2013). With the interpretivism approach, using qualitative research methods we intended to understand the aspects that lead to the biases in AI systems. By taking a data perspective, participants' perceptions of the risks that lead to bias in AI systems were recognized, thus identifying key contributors to biased outcomes.

Myers (2013) argues that social researchers look at the problem from the inside which in this case, is through the social biases that have been created from human perspectives. An interpretive approach will allow us to recognise the different contexts of the situations explored (Myers, 2013). Chen and Hirschheim (2004), point out that much of the research philosophy and methodology comes from the research questions investigated. Our research question aims to answer what the key contributor of bias in AI systems is and where the risks of bias are likely to appear. The origins of bias in AI systems go beyond the positivist paradigm and would not be able to be answered in the needed depth through positivist and quantitative approaches.

Interpretative researchers believe that knowledge is made of the meanings people assign to situations and experiences, in order to make sense of why things are done in a certain way (Klein & Myers 1999; Orlikowski & Baroudi 1991). People's perceptions of the world are shaped depending on their experiences which are often subject to change according to the circumstances. Understanding practitioners' experiences provided us with a great understanding of the biases in AI systems with a clear perspective on data thus highlighting the importance of the research.

3.2 Research approach

For answering our research question, we needed to choose a method for how we wanted to conduct our research. The qualitative approach proved to be a suitable approach for the research as it focuses on understanding the context of a phenomenon by observing people and trying to comprehend their behaviors and experiences through words (Recker, 2013).

When researching our subject, we noticed that there was a gap of knowledge and a request for more research regarding ethical concerns of AI technologies acting biased. Researchers called

for an increased understanding of the role of data in AI as well as the risk of data leading to biased outputs (Mikalef et al., 2022). It was pointed out that there was a need for research on understanding where AI systems go wrong in order to work towards safer systems. Recker (2013), states that a qualitative approach is appropriate for a phenomenon that is not well researched.

Furthermore, in our research, we wanted to get a better understanding of the role of data affecting AI systems as well as where in the process bias takes place. In order to do this, we needed to understand how professionals work and behave with data and how they address issues related to bias. A qualitative approach is appropriate for studying cultural, political or social aspects of technology and is often done by studying data from different data collection methods (Recker, 2013), Therefore, due to the social aspects involving bias in AI, we decided that a qualitative approach would be suitable for investigating our research problem.

Recker (2013) mentions that one should start by analyzing the collected data with the goal of finding patterns, themes and different concepts to further work with. One of the biggest disadvantages of qualitative research is, however, the limitations of generalizing the study to a larger population. Despite this, the approach is still considered to be one of the best approaches for conducting research when the goal is to conduct an in-depth study (Myers, 2013), as we intended to do. However, since our research is based on 8 interviews the generalizability of the study will be limited. With the help of qualitative method research, we were able to increase our understanding of how practitioners work with AI systems, what their views are regarding the possibilities of AI systems acting biased, and find similarities and patterns for further interpretations.

3.2.1 Literature Review Approach

In order to further understand the topic of bias in AI systems, a literature review was conducted before the empirical data was collected. By conducting a literature review we intended to take a more critical approach to AI and comprehend the field of bias in AI systems. The literature review covers foundational concepts for our research such as AI, the data journey in AI systems, and different kinds of biases affecting AI systems. It is intended to understand the process of AI systems in order to identify where biases are taking place. Moreover, real-life examples were taken from both the literature as well as non-academic journals to be able to showcase the relevance of our research.

We used databases such as LubSearch and GoogleScholar when searching for relevant articles. The extensive search led us to associated databases such as ACM Digital Library, IEEE Xplore and arXiv. These databases contained relevant articles from well-known journals. The journals that have been used for this research are mostly from the fields of technology and socio-technological fields, but not limited to these fields. Relevant articles from fields such as communication and medicine were also used. Additionally, most articles used are from scientific and peer-reviewed journals but some non-academic articles have also been used to further point out the importance of the matter that is being studied as well as showcasing some societal issues the subject of bias in AI systems has caused.

The first approach to finding relevant literature was to first conduct a search using keywords such as “Bias”, “Artificial Intelligence”, “Artificial intelligence systems”, “Data”, “Data bias”, “Machine Learning”, “Fairness”, “Ethics”, “Responsible AI” and “Algorithmic bias”. When choosing articles different factors were considered such as the publisher and when it

was published as well as the number of citations. Taking all these factors into account, we intended to gather relevant articles to provide relevant and honest information for our research.

3.3 Data Collection

The data collection technique found to be the most suitable for this study was the set up of semi-structured interviews. The semi-structured interviews facilitate a more flexible approach, which allows for follow up questions (Recker, 2013). Questions were pre-formulated and were able to be shaped throughout the interview to allow for more relevant findings. As researchers, semi-structured interviews provided us with a great balance in executing interviews in a flexible yet controlled format (Walsham, 1995).

As Walsham (1995) points out having a good technique is not enough when conducting interviews, it is also important to possess good communication skills that help the researcher empathize with the participants. For this reason, we chose to ask 'warm-up questions' that made the participants feel comfortable and ready to answer the questions for our research. All following questions asked were relevant to our research as they were formulated based on the findings from our literature review. In addition to our questions, three different real-world cases were presented to the participants to get an increased understanding of their opinions regarding where bias took place in these cases.

The interviews were carried out through video calls since the participants were located in different countries. Additionally, interviews were recorded, giving a full narrative to the questions discussed (Walsham, 1995). However, one could argue that by recording interviews participants might withhold sensitive information (Walsham, 1995). In addition to this, the subject of biases might make participants uncomfortable. As interviewers, we made sure to encourage a safe environment where the participants felt comfortable and were able to share good and high-quality information (Myers, 2013). The interviews were carried out by both researchers in order to have a more critical view. Furthermore, it is important to note that when taking an interpretative approach, it is not only necessary to understand what the participants are saying but also to note their interactions and reactions to the questions (Mero-Jaffe, 2011). Thus the video interview approach facilitated the interpretation of participants' details such as facial expressions and body language. Since participants had different backgrounds, the interviews were conducted in English, however one could argue that a language barrier could result in participants not being able to express themselves with the same fluency as if they were speaking their native language. Nevertheless, we felt that providing participants with the interview guide beforehand allowed for reflections on the main topics they would like to discuss during the interview thus addressing all details needed.

Furthermore, during the interviews, notes were taken to capture the most important details. After the interview was done, interview transcriptions were developed using otter.ai, a safe tool that allows for easy transcription by uploading the recording of the interviews. However due to different accents, these transcriptions were not always accurate, thus the recordings of the interviews were watched at the same time as reading the text to make sure every single word was written properly in the transcripts. Participants' expressions and non-verbal communication were also added to the interview transcript. Once interviews were properly transcribed, transcripts were sent to the participants for approval. Participants were also able to

add, remove or change any other information they found. Below in table 3 we present an outline of the interview details such as date and duration.

Table 3: Outline of the interview details

Participant	Interview Date	Duration	Appendix
P1	Thu, 4/21	39 minutes	Appendix 8
P2	Fri, 4/22	49 minutes	Appendix 9
P3	Sun, 4/24	35 minutes	Appendix 10
P4	Wed, 4/27	76 minutes	Appendix 11
P5	Mon 5/02	27 minutes	Appendix 12
P6	Mon 5/02	66 minutes	Appendix 13
P7	Tue, 5/3	29 minutes	Appendix 14
P8	Fri, 6/5	43 minutes	Appendix 15

3.3.1 Participants

Following Recker's (2013), purposive sampling was done in order to determine relevant participants that possessed knowledge that was valuable to our research. We interviewed participants that had experience working in the field of AI systems, specifically those who had worked with data either in their current role or previously. This approach allowed comparison between people working with data as well as people working on the other side of the model. We felt this would be beneficial to the complexity of our research as it provided us with inputs from different perspectives. Additionally, participants' work experience, as well as their academic knowledge, were important factors when being considered for interviews. All participants had an interest in bias in AI systems. Furthermore, participants from different kinds of organizations were chosen to provide an overview of the way each organization used different kinds of data as well as the organizations' approach to bias in AI systems. Table 4 provides an overview of the participants in our research.

Table 4: Background information about the participants

Participant	Industry	Expertise	Role	Education
P1	Risk Consultancy	AI & Data	Data Scientist	Mathematics & Operation Research
P2	Automotive	Data	Data Scientist	Political Science & Information Systems
P3	Telecommunication	Data	Data Scientist	Statistics, Big data & Artificial Intelligence

P4	Computer software	AI & Data	Team leader Machine Learning Engineer	Computer Science
P5	Risk Consultancy	AI & Data	Data Scientist	Mathematics and Artificial Intelligence
P6	Computer software	AI & Data	Senior Artificial Intelligence Analyst	Artificial intelligence and Machine Learning
P7	Risk Consultancy	AI	Manager of digital ethics	Technology policy
P8	Computer software	AI	Head of Machine Learning	Mechanical Engineering

The participants of our research were identified and initially contacted via LinkedIn, receiving a short introduction about our research. Once they showed interest in participating, we sent them more detailed information via email. Furthermore, some participants requested additional details through calls. Once our participants agreed to be interviewed, a date was set, and the interview guide was sent in advance. One of the reasons the interview guide was sent was to increase participants' confidence and conform as English was not the main language of most of the participants. Furthermore, this practice allowed participants to read in advance for the main questions thus increasing the probability of obtaining well-thought answers.

3.3.2 Interview guide

Interviews were conducted by following an interview guide. An interview guide in a semi-structured interview acts as a half-completed script, where questions were written but there was also room for follow up questions during the interview (Myers & Newman, 2007). The interview guide was developed following the same structure as the literature review. Firstly, it started with an introduction of ourselves and the research as well as outlining participants' rights and asking for permission to record. Followed by warm-up questions where we could get to know more about the participants' experiences and education and create a more conversational setting. Subsequently, questions that were relevant for our research were developed taking the main concepts identified in the literature review into account, focusing on firstly the topic of AI, continuing with the topic of data and finally connecting both with the bias perspective. To conclude the interviews, we decided to have a question on team diversity, as it is mentioned in the research as a good way to detect bias. Once the questions were answered, the cases were introduced. In conjunction with this, a question relevant to all the cases was presented. Each of the cases consisted of real-life events where bias had occurred. The concepts and relevant questions and case examples can be found in table 5, the full interview guide can be found on Appendix 1.

Table 5 : Interview guide

Concept	Question
Warm-up questions	- Please provide us with a quick introduction about yourself.
	- Can you tell us about your experience with data and artificial intelligence?
Artificial Intelligence perspective	- There are many different ways of defining AI, we define it as: *"AI is the ability of a system to identify, interpret, make inferences, and learn from data to achieve predetermined organizational and societal goals*." Do you agree with this definition? Would you add/change/remove something?
	- The literature that we have read points out that there is a need for addressing the "dark side" of AI such as the risk of AI systems acting biased. Is this something that is talked about at your workplace and among colleagues?
Data perspective	- When collecting data, do you have specific requirements that you follow or other aspects you try to consider?
	- How do you make sure that the data that you use is of high quality?
	- Can you describe the process from when you get access to the data you are working with until it is used as input for algorithms?
	- Can you ensure there is no risk of bias in your data, algorithms or the systems you work with?
	- Do you work in a more precautious way in relation to bias when working with data that contains demographic information?
Bias perspective	- How do you work in order to detect biases?
	- Is your organization able to identify bias, and if so, what type of biases have your organization identified?
	- Based on your previous experiences, what is the most common type of bias?
Team diversity	- How diverse are your teams? Do you think having diverse teams can have a positive effect on detecting bias?
Cases	- Starting from data collection to the model deployment of the AI system, where would you say is the highest risk of bias occurring?

The interview guide enhanced our interactions with the participants and helped us make the most out of the interview. As interviewers, we made sure that all relevant questions were

covered. At the end of the interview, participants were asked to provide any information they felt could be relevant to the research. One of the participants provided us with recommendations for people to interview, allowing us to increase our number of participants. This technique is called snowballing as one participant can lead to many others and those other participants could lead to more (Myers & Newman, 2007).

3.3.3 Bias in real-life cases

Once the questions were answered, the cases were introduced together with a relevant question. Each of the cases consisted of real-life events where biases had occurred in AI systems. Participants needed to identify the origin of the bias and the source for each case as well as motivate their choice. The cases were displayed through a presentation via Canva, which is an online platform used to create visuals. The purpose of the cases was to obtain relevant insights from the participants on the different origins and sources of bias in AI systems. Firstly we presented an illustration of the different stages in the AI process, created following Suresh and Guttag's (2021) approach to identifying bias.

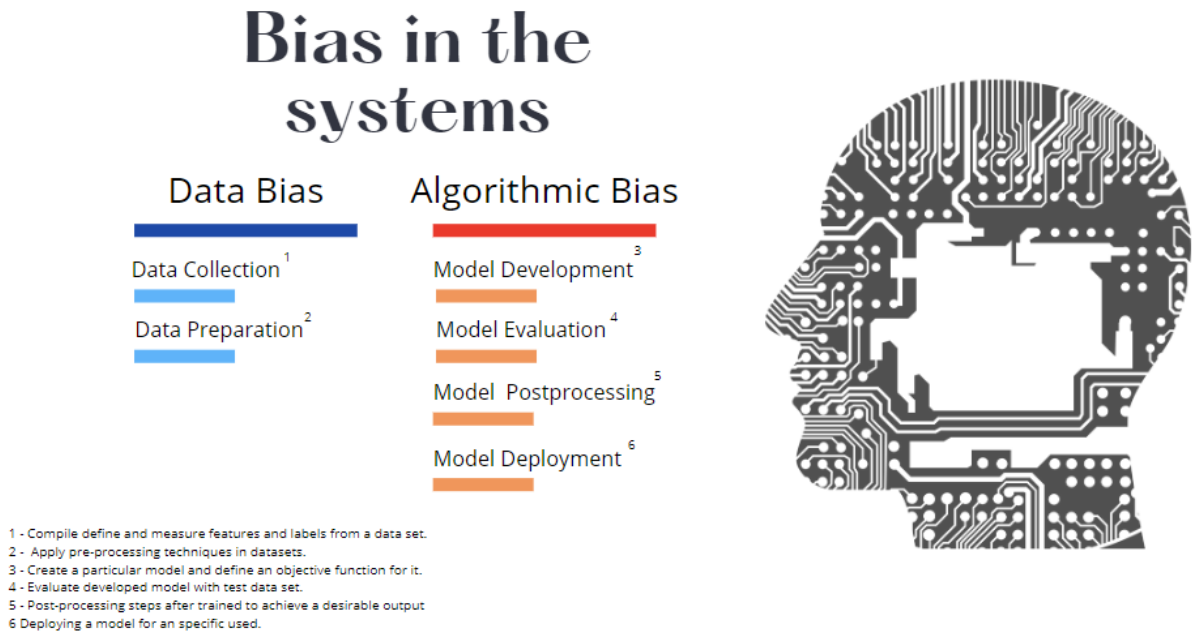


Figure 2: Stages in data and algorithmic bias presented to participants

When choosing the cases we wanted to include different kinds of cases that were relevant to our research. Therefore, we decided to include our first case (Appendix 2 & 3) where an AI system was used for the automation of passport renewal. The second case (Appendix 4 & 5) was a system that used AI to predict the risk of someone re-offending. The last case was replaced after conducting the first interview due to the fact that we felt that it was more similar to the first case than desired. Therefore the official last case (Appendix 6 & 7) presented an incident where job ads within the STEM-field were more frequently displayed to men.

Once each case was explained and the participants provided their answers, an extra piece of information regarding the case was presented, followed by asking the participants if they would like to change their answers based on the newly obtained information. This experiment was carried out to emphasize the complexity of biases in AI systems. By taking cases from real life we also highlighted the relevance of the issue. This part of the interview was highly praised by participants as they had the opportunity to motivate and explain their points of view with examples.

3.4 Data analysis methods

When the data has been gathered it is essential for the findings to be analyzed and interpreted. Our research was exploratory rather than theory-testing, therefore, the data analysis was done through a bottom-up approach as recommended (Myers, 2013). Memoing was also done right after the interviews to capture the feelings and thoughts that were experienced when conducting the interviews. Recker (2013), states that procedural memoing is an effective way to summarize how the interview was, if something special occurred, how it was perceived and the atmosphere of the interview. He mentions that researchers also can include some initial ideas and interpretations in their memos for further analysis which can be used as a guide for the coding process.

There are several different ways for analyzing qualitative data. After closely evaluating the different approaches, we believed coding to be the most suitable technique for our research. Open coding was found to be the most fitting technique for our research. Using open coding eased the process of identifying key concepts for the practitioners' way of thinking. By using coding for our analysis, we were able to categorize the large amount of data that was obtained. Coding is the perfect time for realizing if data is missing and if the researchers need to gather more data to better understand concepts and explanations (Recker, 2013), this gave us the opportunity to see if we needed to reach out to our participants for follow-up questions.

The main concepts were colored with completely different colors. The main concepts were then divided into subcategories which were colored in the same color but in lighter shades.

Table 6 Code used for the findings

Concept	Color	Aspect
Artificial Intelligence		AI dark side
		AI definition
Data Quality		
Bias		Data Bias
		Algorithmic Bias
		Community Bias
Societal effects		Team Diversity
		Awareness

3.5 Ethical considerations

In order to protect ourselves, our university as well as the participants of our study, it is important to take ethics into consideration when carrying out research (Myers, 2013). Myers (2013) defines ethics in research as taking moral principles into account, throughout all aspects of our study from the research design, and data collection to analyzing and writing the findings of the study. As researchers, we made sure the below ethical considerations were followed.

Since interviews were the chosen technique for data collection in this research, we firstly provided participants with a description of what was being researched and what would be done with the results, followed by a description of their ethical rights presented in the interview guide. Participants were aware of their rights and knew they could withdraw from the interview at any given time, as participation was voluntary. Similarly, following Patton's (2015) guidelines, we obtained informed consent from all participants before starting the interviews.

As mentioned before, interviews were carried out via video call and were later transcribed. Before starting the interviews, participants consented to being recorded. According to Myers (2013) and Recker (2013), it is important for researchers to make sure the privacy and confidentiality of the participants are preserved when recording and transcribing the interviews. In order to protect participants' privacy, names and company names were not included when transcribing the interviews. The transcripts included detailed information about the interview, including facial expressions, body language and involuntary sounds to be able to analyze verbal and non-verbal communication (Mero-Jaffe, 2011). Moreover, participants got a copy of the transcribed interview, which gave them the opportunity to confirm their answers, correct errors and add more information if needed to enhance the quality of the data (Mero-Jaffe, 2011).

Ethics does not only affect the data collection section of the research, thus ethical aspects were also taken into account when analyzing and writing the research. For instance, in the analysis phase, we ensured data was properly represented in an honest manner (Myers, 2013). Recker (2013) mentions that researchers have an obligation to explain the techniques used to analyze and report the data, as we are explaining in the research design. Additionally, Recker (2013) points out that the evaluation of the ideas should be impartial, fair and with the use of complete data, thus we had taken responsibility for our findings regardless of the results. Similarly, when writing the findings Recker (2013) also points out that researchers should use appropriate language to avoid any kind of discrimination, since the topic of biases addresses discrimination, we made sure that it was being mentioned in a sensitive and cautious manner. Furthermore, to avoid plagiarism, full acknowledgement of people's work is written in the thesis. With that being said, one of the last ethical issues we addressed was to make sure our research was ready to be published.

3.6 Scientific quality

One way to evaluate the scientific quality is by the psychometric properties. The psychometric properties are divided into two parts, reliability and validity (Bhattacherjee, 2012). These can in turn be split as internal reliability, external reliability, internal validity and external validity (LeCompte & Goetz, 1982). The adequacy and accuracy of our methods for

measurements can be evaluated with the help of these (Recker, 2013). Easily said, reliability describes how reliable something is, meaning that if several studies, independent of each other were conducted and they all reach the same results the reliability can be considered high.

External reliability refers to whether or not other researchers will be able to replicate the study with the same outcome. Internal reliability is focused on how well the involved researchers agree on how the data should be analyzed (LeCompte & Goetz, 1982). By using a coding system that both researchers followed we tried to increase the internal reliability of this study by providing a detailed guide on how the coding of the empirical data was conducted. As earlier mentioned, the data collection for this research was conducted with the help of interviews. Maintaining high external reliability is often hard when conducting qualitative research and interviews. The reason for this is that it is hard to recreate the exact same setting that once was created for the interviews (LeCompte & Goetz, 1982). Furthermore, the use of semi-structured interviews makes it even more challenging. Meaning that despite there being an interview guide, the interviews may turn out very different which will result in the study and its results being even harder to replicate. By providing a detailed description of how the study is conducted, what questions were asked and how we worked with the empirical data we are hoping to be able to slightly increase the external reliability of our research.

The validity, on the other hand, is about how well the collected data measures what it is intended to measure. External validity aims to investigate if the results of the study are generalizable (Recker 2013) and internal validity focuses on how well the interpretations of the collected data are aligned with the stories the participants wanted to tell (Bryman, 2016). Bhattacherjee (2012) argues that the best way to measure validity is to combine a theoretical approach with an empirical approach. Therefore, we have conducted a literature review that laid the ground for our interview guide that was used when collecting empirical data from our interviews. By combining these two approaches we worked with two different data sources and had the opportunity to compare them in order to draw relevant conclusions and increase the internal validity of our study.

Furthermore, the study was based on 8 interviews and therefore the external validity will be limited. Patton, (2015) and Recker (2013) consistently argue that there is a consensus regarding quantitative methods being counted as a better approach for generalization than the qualitative approach. This can lead to our findings being limited in terms of generalization. By interviewing practitioners from different companies and countries we aimed to slightly increase the external validity by including the understanding and knowledge of 8 different people with different backgrounds. This led us to obtain their rich and detailed views. By using semi-structured interviews, we were leaving room for the participants to add comments to make sure their stories were perceived in the way that they intended. By doing this we eliminated the risk of misunderstandings.

4 Findings

The following chapter provides the key findings of our empirical data analysis. When doing the analysis, the findings that were relevant for answering the research question have been gathered and will be presented below. Some participants provided us with extensive answers that were not relevant to our research, thus those answers have been shortened and the key takeaways have been included in this section resulting in the warm-up questions aiming to get to know our participants not being included.

Additionally, it is worth noting that not all the participants were asked exactly the same questions, because of the decision to conduct semi-structured interviews, therefore some subtitles of our findings may not include all participants. The participants' identities have been anonymized as well as their workplace, therefore the participants will be referred to as P1-P8. Furthermore, one of the cases was removed and replaced with another one after the first interview.

4.1 Defining AI

The following definition of AI was provided to the participants *"AI is the ability of a system to identify, interpret, make inferences, and learn from data to achieve predetermined organizational and societal goals*." All participants, with the exception of P6(L10) and P7(L8), agreed with the definition. They also agreeingly stated that AI is hard to define and could mean many different things in different situations and circumstances. P1(L6), P3(L8) and P4(L8) completely agreed while P2(L9) and P5(L10) partly agreed and shared their reflections. P2(L9) considered AI to be an umbrella term and continued by saying that it is hard to pinpoint what AI is as its meaning is subject to change and finding patterns in data would also make a good explanation of the term. P4(L10) adds to that by arguing that what used to be referred to as AI does not necessarily have to be considered AI today.

When the definition was presented, P5(L10) mostly agreed, adding an emphasis on it being the system's ability to perform tasks humans usually do in order to make our lives easier and automate processes. P8(L9), just like P5(L10) agrees with most of the definition but emphasizes that different people probably would define it in different ways resulting in people with technical knowledge focusing less on the part of the definition that talks about achieving predetermined goals while people with less technical backgrounds have a tendency of focusing more on that. Lastly, it's mentioned that despite P8 having a technical background, the participant tries to integrate the social parts.

P7(L8) provided us with a different explanation introducing us to thinking of AI as descriptive and prescriptive systems. The predictive systems predict different things such as how likely it is that a client will commit fraud and the descriptive system would instead pay more attention to existing data and find patterns without making predictions for the future. Despite completely agreeing with the provided definition, P3(L8) made some similar reflections as P7(L8) regarding interpretation and inferences, pointing out that some situations are more about finding patterns rather than predicting solutions. Lastly, P6(L10) stated that the definition that was provided during the interview was more toward Machine Learning and that AI is

a broader term that needs to have a broader definition. Further, P6(L10) states that their preference is the definition that is provided by Russell and Norvig in his famous book.

Table 7: Definitions of AI – Findings

Participant	Agrees with the presented definition of AI	Comments from the participants
P1	Yes	Mentions that it Is hard to define since a variety of things could be counted as AI, but a more societal definition makes sense.
P2	Yes, partly	Argues that it is hard to pinpoint what it is. Some things that used to be referred to as AI does not have to be AI today. Believes AI is an umbrella term for many different things such as robotics, IS and statistics.
P3	Yes	Mentions that the definition changes depending on the circumstances, especially the interpreting and making inferences part.
P4	Yes	Refers to it as a moving target, that pushes limits and could be more of a philosophical debate. Defining AI could be quite challenging to define.
P5	Mostly agree	Argues that it is mostly related to computers and their ability to perform tasks that humans can do to make life easier and automate tasks.
P6	Disagrees	Argues that this definition is more suited for Machine Learning rather than AI. Explains the definition of AI as broader, stating that this is too specific. Would prefer Russel and Norvig's definition provided in his book which states that AI is intelligence demonstrated by machines.
P7	Provides their own definition	Makes a division between descriptive and prescriptive systems of AI.
P8	Yes, partly	Argues that the definition varies depending on who you talk to, if a person is technical or not some adjustments might be done. Technical people have less focus on the predetermined goals part of the definition.

4.2 The dark side of AI

All participants were aware that AI systems have a dark side, where discriminatory results can be produced. However, when asked if this is a topic that is being addressed at their workplace half of the participants partially agreed. P1(L9) and P5(L12) point out that it is mostly talked about in the context of regulatory compliance. P6(L22) mentions that there is a concern

regarding the topic but it is not something that is talked about often as it is a recent topic. However, P2(L13) P7(L10) and P8(L13) completely agreed on the dark side being an important part of their current role and company. P2(L13) and P8(L13) further reflect on how this topic was not as important in their previous company. Similarly, P4(L13) points out that it is not a priority topic in their workplace. Nevertheless, P4(L13) takes the dark side of AI seriously when doing their job as they have studied this topic before. P3(L11, L13) notes that the topic is not addressed in their company, however, is being talked about amongst colleagues from a philosophical perspective.

The way that the dark side of AI is being addressed varies. P1(L9) and P2(L13) observe that model validation plays an important role in addressing the dark side of AI, P1(L9) makes special reference to the design of a trustworthy and transparent AI system. Furthermore, P7(L10), describes that the whole model lifecycle should be looked into, mentioning that this is the best way of understanding the different stages that can introduce bias into the systems.

Table 8: Dark side of AI – Findings

Participant	Is the dark side of AI raised at work?	Comments from the participants
P1	Partly	Is being talked about in a regulatory way. Mentioning that developers are often already busy with different tasks, thus they need external support to comply with regulations.
P2	Yes	Frequently spoken about, their mission is related to the dark side of AI working with self-driving cars, highlights that it was different in previous workplaces.
P3	No	Mostly with colleagues, focusing on psychological perspective but not with work.
P4	Partly	Depends on the environment, the views are different in the academia and the industry, P4 reflects that if you have previous knowledge on the subject you are more likely to take it into account.
P5	Partly	Partly but only in terms of GDPR, should be talked about more from the effects of the models that are created, not only in the context of GDPR.
P6	Partly	Is a recent topic, with growing concern, however, currently, this is not being talked about at their work because the data they use is not as sensitive.
P7	Yes	Most common way that it is talked about is by looking at the entire model lifecycle. Ideation, building the model, testing the model, training it and putting it out for production.

P8	Yes	The current company looks at it as an important part of the process and machine learning models. However, the last company was not as focused on the dark side because it involved another kind of data.

4.3 Data in AI systems

It is evident that companies work in different ways with data collection. Some companies such as the workplaces for P1(L15), P2(L16), P5(L22), P6(L28) and P8(L21) receive their data from external resources such as other companies or open-source data sets and do the quality checks for them. P3(L15) mentions that their data is generated by their customer. P2(L15) makes an interesting point arguing that many organizations don't really pay attention to data quality, the focus is to implement AI because it is cool rather than focusing on using it in a responsible way.

In terms of data quality, the approaches are different amongst the participants. P1(L19) describes the initial stage as taking a look at the data and evaluating it, asking themselves questions to see if anyone is singled out or overrepresented. The next step is to look at possible proxy variables that could cause problems. P1(L19) gives an example of how the postal code of a neighborhood could give indications of people's ethnicity. If there are data points that can be misleading it is important that these are not reflected in the models. P4(L28) mentions that even if a sensitive attribute is ignored, it can always enter the model in a different way, but continues by stating that when using demographic data one should be extra careful. Continuously, P1(L30) argues that it is important to be observant with all data sets since no data sets have exactly the same flaws.

P2(L16) mostly works with data that is collected from outside their organization. Even though they are not responsible for the collection part, they do hold the responsibility of making sure it maintains high quality. P2(L16) and P7(L16) mention data labeling as one action for ensuring high data quality and P2(L20) continue by saying that the data scientist is often powerless, receiving the data and during the process realizing that the data is wrong or bad. Moreover, P2(L18) states that checking your data pipelines for assertions in the code and the database is also helpful for maintaining high data quality. Sanity checks are also mentioned by P2(L18) to be important but quite challenging in some cases because there is a need for people that possess the relevant knowledge and have good communication with the data team.

> Sometimes you will think that okay, but of course, no one is shopping more than 60 times in a month, right? That was not reasonable. But actually, if you're two people in your household, and you always buy like a coffee on the way to work at your local market, because you live close to one so we can like is first we get a number we like, unreasonable. And then you can quite easily actually come up with a scenario where other data actually is reasonable. – (P2:L18).

P3(L19) mentions similar statements, saying that they do such checks as well, a person being 200 years old would not be reasonable and would be an indication that the data is bad. P2(L18) continuous with another example:

> …if you are buying different apples, you might just notice that when they have the same price, I can put them in the same bag. Right? Yeah, that will give you low quality data on the both apples because it will think that someone only bought, you know, Pink Lady apple, but actually had a collection of different. And then there's a challenging getting the store people to understand like, that's actually not the store because sometimes it's what the person in the cashier **machine sounds** into the system. But sometimes it will be like, just you know, how, how do you make it so that your customers won't behave in a way that creates bad data. - P2(L18)

P2(L41) points out that now that the work involves data on different traffic situations it is important to collect data from different countries. Data on traffic situations in Europe might not be suitable for other parts of the world. P4(L17) shares a similar belief of the subject stating that it does not have to be about the amount of available data, it is rather about the data having an even distribution making an accurate representation of the society or the place the system is intended to be used in. Therefore it is important to know the project to make sure that relevant and well-represented data is being used. P4(L19) continued by stating that a fair data distribution and data representation are important for ensuring high-quality data which is also confirmed by P3(L19) and to be able to have this P4(L33) once again emphasizes that it is important to know your project. Additionally, P4(L17) believes that validation of the results with real-world data is also important and will indicate if the data is maintaining high quality. P5(L34) confirms this stating that they also consider relevant data for projects to be essential when making sure the data is of high quality, giving the example of a system for car damages that needs to be trained on car pictures. The representation of the different car types in the data set is important and should be of the same distribution as the place the system is being used. People in countries outside of the European Union might drive different cars and that should be reflected in the data set.

Moreover, P5(L20) adds that the preprocessing step is important for making sure there are no faulty data or missing data points in the data set. If it turns out that there is some missing data there need to be guidelines on how to manage that to get high-quality data. P5(L20) mentions some possible solutions being deleting the data points, using an average or just simply trying to find the right value. P5(L30) adds that when working with demographic data the preprocessing is even more important to minimize the risk of bias occurring and they often require the clients to not include sensitive data, such as people or license plates in their pictures. Lastly, P7(L16) emphasizes the importance of data labeling for high-quality data and avoidance of biased data arguing that it is an entire profession today, where people spend their entire days labeling data.

The analysis is also considered to be an important part of maintaining high-quality data according to P6(L28), who continues by saying that when they work, they are extra observant when checking the variety of the data. Since their work is focused on chatbots it is important to have a large variety of data to train the model with so that the chatbot is able to respond to all different types of messages. When asked about how they define a good data set the participant highlights that it is a hard question to answer but in the case of chatbots variety has the most impact. It is important that their data set contains many different ways for expressing the same thing, therefore high-quality data in this case means a data set with high variety (P6L34). Further, P6(L34) adds that ”*…it's better has a small data set with your high variety of the data, instead of to have a huge data set with much more equals messages”.* P4(L17) agrees by explaining that there is limited use in having a really big data set that contains

samples that have similar features, therefore, data diversity is very important. Having data to further train your model and make it better as it launches in new countries is also pointed out as important by P6(L46):

> So right now, for example, we add a French company in our system. And we are not we are now collect data for French people to try to fine tune the machinery models for French as well. So it's really important and we can clearly see the performance in going down in we try to generalize the problem we have to work with data for all countries. And one important thing as well is the balance of data. So I cannot have a lot of data from Brazil only and a few samples for from other countries just doesn't make a good model for this. So it's important to have a lot of data from different ways. - P6(L46)

Lastly, P6(L36) says that they do a quality check when the laboratory phase is over and the product is put into production. During the quality check, one can see if anything is wrong, such as classes missing. The last participant, P8(L6) was currently focused on developing Machine Learning models to be used for bias detection, therefore, they did not aim for high-quality data but rather data that gives a correct representation of the real world regardless of attributes that can be classified as being racist, sexist, ageist (P8:L19).

4.4 Bias in AI systems

As evident in appendix 16, Most of the participants agreed that there is no way to 100% ensure that data, algorithms and systems will not contain bias (P1:L23; P2:L30; P3:L25; P4:L26; P5:L26 & P6:L44). However, P8(L46) points out that depending on the system being used, a bias free model could be achieved if the model is designed in the right way. Nevertheless, P1(L23) and P6(L44) remark that it is very hard to avoid having biases in the systems, with P2(L30) and P4(L26) arguing that even though biases are not able to be evaded, that does not mean that there is no way to reduce them. P4(L26) puts emphasis on awareness of there being biases that can let one decide on acting on the bias. P1(L23) agrees, mentioning it is a matter of knowing if the bias is intended or not. P5(L28) continues by saying that despite a system being free from bias not being something that one can promise there are ways of mitigating the risk of it as well as detecting bias which in some cases can be done using algorithms, which is confirmed by P6(L61), P7(L21) and P8(L4).

4.4.1 Bias from a data perspective

P4(L26) argues that how bias is defined matters as well as being aware of the bias so one can pursue the process of tackling the bias. Additionally, there is a mutual agreement regarding data being the main cause of bias in AI systems with P1(L34) saying bias is *''... if not completely, then very close to a data problem. The models are just like learning whatever you show them''* which is also confirmed by P3(L35) stating, *"I believe that is it's the most bias types come from data''*

P1(L21) has limited experience working with data sets other than open-source data sets where the bias is normally already known. The participant provided an example of a data set that contained a class with only one data point for Eskimos leading the system to deny the Eskimo a loan since the decision was being based on that single data point. Imbalanced data sets are mentioned by P4(L35) as yet another reason for data bias. Using the step of anomaly

detection one can detect things such as outliers that deviate from the data set (P4:L35). Furthermore, P7(L16) says that the step of creating data sets for training the models is a risky step with a risk of bias gradually developing. In addition to this, P7(L16) pays special attention to the data labeling used for the algorithm, mentioning that if it is not performed properly it can lead to bias since things can have different meanings depending on the interpretations of the person doing the labeling:

> …data labeling, that's where you see a lot of bias creep in, when training data for an algorithm is created, that the algorithm is trained on the model is trained on that people,usually those are like AI, people that studied AI,....their only job the whole day is labeling data… and there you see that you can quickly get a kind of bias, if you have only have white men labeling data, you are going to get a bias in the outcomes – (P7:L16)

P3(L33) states that using data regarding attributes such as political views and religion can be quite dangerous and result in biased outcomes. They don't collect this type of data, but do not have any protocols or guidelines for how to detect possible bias. P3(L33) argues that some sensitive attributes should not be included in the data and model. However, P2(L32) and P4(L28) disagree saying it depends on the case, with P2(L32) saying *''And it's not like remove all of this sensitive information, and then you'll be fine. Absolutely not''.* Additionally, P2(L22) also argues that a careless approach is to do the opposite. Instead of reducing variables, some people put all the data into the system without curating it thus ignoring possible correlations.

P2(L30) and P4(L19) talk about the importance of having different attributes in the data, mentioning that just removing a variable does not necessarily remove the bias from entering the systems as the correlation between variables matters. P2(L30) then gives the example of ice cream and sexual preferences versus pads and gender.

> I cannot check is there a correlation between sexual orientation and ice cream preference? I cannot check. I will not know if the ice cream like preference actually represents sexual orientation information, the way I can with sanitary pads and like menstrual pads and gender that I can check. – (P2:L30)

Just like P2(L30), P1(L19) mentions that certain combinations of attributes can lead to unintended bias, saying that the combination of a certain nationality and gender could exhibit bias without anyone guessing the outcome will be biased. P1(L19) also mentions connections between different attributes, calling them ''inherent'' in the data sets, giving the example of ethnicity embedded in the postal code of zones with higher contraction of dark-skinned population.

> So I think, in our case, the demo shows is that, for example, ethnicity, you have something like postal code that can be quite, you have certain areas with ghettos or something where the ethnicity is, to some part reflected in the postal code of the address. And then, so even if you just like get rid of ethnicity as a feature, the bias is still inherent in the data set – (P1:L19)

4.4.2 Data bias affecting algorithms

All participants reiterated the importance of the training data used in the algorithms (P5:L18 & P1:L21) as algorithms learn from data (P5:L58; P7:L16 & P3:L31). P5(L36) mentions that if bias were to be found in the data, it is necessary to fix it to avoid a biased model. P3(L27) agrees, mentioning that there is a need to ensure data is not biased, later adding that an algorithmic perspective should also be taken into account. In addition to that point, P1(L65), P2(L68) and P4(L52) argue that any bias found in the data should be taken into consideration when developing models. For example, P2(L68) mentions that historical data favors men over women and specific models should work with this thought in mind thus being able to mitigate the bias found in the historical data.

> It's kind of like data generation, you could almost say. But you could probably have the same thought I had there about maybe that's what we should call algorithm bias and that you have not kind of taken that into account when you try to develop your model. – (P2:L68)

P1(L65), P5(L42), P6(L61), and P8(L79) observe that if any bias were to be found in the data, the model should be able to identify it. P4(L52) agrees by saying that the quality of the data is measured in the model, thus allowing for the removal of bias if they were to be found. P8(L79) adds to that by saying the models can be tested by using different types of data, giving emphasis to adversarial testing, thus if bias were present in the training data, the model should notice that something is wrong. P1(L65) emphasizes that a lack of attention in the model process using biased outcomes can lead to biased outcomes

> The models are not inherently biased. They are like just mathematical techniques to model data. So if the bias has to come from somewhere, in my opinion, it's on the data side or like a lack of attention during your model process. Because then bound data is bound to be biased to some extent. During the modeling process, that's where you would then see. – (P1:L65)

4.4.3 Bias from an algorithm perspective

P7(L10) takes on a more algorithmic approach to bias, arguing that the topic of training data has been the subject to a lot of discussions. P7(L10) emphasizes the importance of paying attention to the different stages of the model life cycle instead:

> So from ideation, to Yeah, the building the model, then testing the model, training the model, putting the model into production, and then even the end, like retiring a model, all those different stages can introduce bias. - (P7:L10)

When mentioning model development, P1(L47), P2(L68) discussed that different kinds of biases such as the representation of historical bias can be tackled in the developing phase, otherwise these biases can be amplified by the systems. P3(L56) remarks that since models are designed and manipulated for desired outcomes, this can affect the processing of the data presented, leading to biased outcomes. Similarly, when referring to Case 3, P7(L45) notes that fairness metrics were used in the systems thus favoring some and discriminating against others. P2(L30) continues by mentioning that the necessary variables should be taken into account when building the model. P2(L30) proceeds to give the example of deciding on gym grades in gym class, if gender is not taken into account females would perform badly:

> If you try to like build a model using maybe the strength of the person, so how much they can do in like, some kind of workout like… But then probably the strength overall would be a pretty good indication of what grade you have, like probably people have higher grades generally also are stronger and you will find a positive correlation there. And you can kind of build a model with it. But if you were not to include gender as a variable, then all girls would get lower grade, and the guys would get higher grades. Because they're generally stronger. And if it's a positive model is just supposed to the relationship between them. So, then you actually want to include gender to ensure the fairness. – (P2:L30)

Following model evaluation, P2(L58) and P5(L66) mention that this step is tied to data collection. P6(L61) adds that the data preparation can be checked in the model evaluation. Similarly, P8(L79) mentions that adversarial testing can be done to make sure there is nothing wrong with the system. P2(L39) suggests that understanding the output of the model is important to make sure it is performing properly in different subgroups.

Finally, model deployment was mentioned by P4(L33) and P7(L21), both participants touched on the examples of financial systems that were biased depending on the locations of their deployment. Additionally, P1(L11) concludes that models are not often developed with malicious intentions. P8(L79) agrees with this statement by proving an example where models are being compared to an innocent child

> The model, the algorithm actually, inherently is unbiased when you start, it's what you use to train it, the teacher, that it's like an innocent child, that's our model. If you raise your child telling him that, black people are bad, they will think they're bad. But if you don't tell that to your child, and you just let your child meet different people and appreciate them, then the output of that child or model would be that these people are not bad. - (P8:L79)

4.4.4 Bias from a human perspective

P5(L42), P3(L11), P2(L62) and P4(L52) share the thought that bias is a societal issue rather than a technical issue. P5(L22) and P3(L35) accentuate historical bias as being one of the most common types of biases that contributes to data and algorithmic bias. P5(L40) criticizes the history of humanity, '*You could say, we were biased in the past. So now all the data is biased."* and argues that it can cause consequences in data thus affecting the algorithm.

Furthermore P3(L11) mentions that *"people are broken''* and AI learns from people's behaviors, making it impossible to ensure there is no bias in the systems. In addition to this approach, P2(L62) argues that the issue can become more philosophical, even though we might have a clean data set, there will still be a portion of bias as these biases are still existing in our everyday life. Likewise, P4(L52) mentions that the data is only representing reality, which is already unfair and unbalanced.

> Maybe you have perfect data, but there's still bias because there's bias in the world. So that's, to some extent, this becomes a philosophical discussion about to what extent data is a good representation of the world.- – (P2:L62)

When talking about the workplace, P4(L22) mentions that the lack of understanding of the whole process leads to misconceptions from different teams, which can have an impact on the bias. P2(L28) believes that there is a possible risk of employees, such as data scientists, transferring their own bias into the systems, giving the example of predicting if someone has a child based on their age. In order to mitigate some of the biases a diverse team is necessary. All participants agreed that a diverse team was an important factor in detecting bias in the systems (Appendix 17). P1(L36) and P2(L47) agree on diversity as a provider for a deeper understanding of the different perspectives of bias. P3(L39) and P5(L44) adds to that by noticing that diversity in the teams can provide different insights on predictions from the data. P7(L16) reflects on diversity from the data labeling perspective, by mentioning that labeling is tricky thus a diverse team might help combat discrimination.

P8(L65) emphasizes that diversity helps broaden perspectives, adding that people with skills in both soft and hard science should be included in the teams. P7(L16) mentions that diversity should not only come from culture and ethnicity but also from academic backgrounds. P8(L60) demonstrates this argument by mentioning that their team deals with researchers that help review their data as well as stay in the loop of the latest research regarding bias. P2(L47) recalls that earlier on, there were a lot of issues due to the lack of diversity in the field of information technology, as for example some of the staff in their last team, were not as attentive to noticing if the systems were discriminative or not. However, one important view to consider was the one of P4(L39), which discusses that diversity doesn't necessarily mean having the capabilities to detect bias, paying special attention to the origin of the problem or if is a technical or a more human-related issue:

> Because so for example, if it's a very technical problem, like cell phone, fault detection, it might not be the rest of the team might not be that relevant. But when it when you will, closer to some more human oriented problems. So as fraud or advertisements. - (P4:L39)

4.4.5 Most common type of bias in AI systems

Most of the participants (P1:L32; P2:L45; P3:L35; P4:L33; P6:L59 & P8:L60) agree that most data bias is due to the sampling or representation being bad. P3(L35) adds that historical data bias, reflecting the biases that are embedded in our society, is also a common reason, which is also confirmed by P5(L40). Further, P3(L35) adds that even algorithmic bias plays a role, however, the effects are limited but emphasizes that all three (representation bias, historical bias and algorithmic bias) are important. Based on P6(L59) previous work experience most of the biases have been related specifically to the data collection parts and states that unfortunately, that is also the hardest part to identify bias in.

Despite stating that representation bias is the most common type of bias, P2(L45) argues that it is however dependent on the specific project, for some types of examples some specific types of biases may be more common while other projects more often struggle with other types of biases. The participant gives grocery shopping as an example. The data is collected from the customers' previous purchases, resulting in fewer datapoints for customers who shop less frequently. P6(L49) adds to this, by mentioning that they will have more data from customers from the most popular cities, rather than the ones from smaller cities. Furthermore, P8(L60) reflects that even though the company tries to remove as much bias as possible, there is still the problem of representation of certain communities, as there is not enough data on these minorities.

As previously stated, P4(L33) agrees with the representational bias being the most common cause of bias and continues to mention imbalanced data sets as problematic. P5(L40) shares reflections regarding the bias that is and has been embedded in our society for centuries, which resulted in data sets that are affected by historical bias.

An unanticipated answer was received by P7(L26) who broke the pattern of saying that the most common type of bias is neither historical nor representative. P7(L26) shared some thoughts on how people always tend to think that technology is always the right solution, but making this assumption is biased according to P7(L26).

Table 9: Most common type of bias

Participant	Common type	Comments from the participants
P1	Representation bias	Agreed representational bias is the most common type of bias providing the example of computer vision where white males are predominantly leading to underrepresentation of women, especially minorities. Emphasizing that the algorithms will perform better on a male from Scandinavia.
P2	Representation bias	Argues that it is dependent on the case and the main problem one is trying to solve. The participant argues that representation is often occurring but also argues that many different things can go under that name. Gives the example of grocery stores that collect data about their customers. Some customers shop more than others, resulting in more data being collected in regards to the ones that shop more frequently.
P3	Historical bias Representation bias. Algorithm bias a bit too	Emphasizes that all of them are important. Says that the data that is used comes from our history. Points out sampling as important as well as algorithmic but says that the latter has small effects.
P4	Representation bias	Many of today's problems are due to imbalanced data and anomalies in the data set. Some of the most common errors are by the dominance of a specific attribute, where data set representation plays an important role.
P5	Historical bias	Speaking from personal opinions and experiences, P5 argues that the main issue is the bias in our society that is now reflected in our data, hence historical bias.
P6	Representation bias	Mentions it is hard to identify in data. They may collect data in one place and then the model is used in a different country or region.

P7	Bias towards the benefit of using AI	Argues that saying that digital solutions are always the best solution is a biased statement.
P8	Representation bias	Mentions that the challenge is representation, for example currently their model has a lack of representation of specific neurodiversity and physical disabilities (L60). This is due to the reason that there is a lack of understanding on how these communities perceive systems differently. P8 emphasize the even though they are trying to fight on bias in every way possible, some groups might be underrepresented due to the lack of information there is on those groups.

4.4.6 Bias detection in AI systems.

Participants discussed different ways to detect bias, with only P3(L33) mentioning they do not work to detect bias in their systems, instead, they avoid using sensitive data that could cause discrimination such as religion or political views. P8(L27) provides a detailed answer by breaking it into 3 different parts, data, model and system. Adding that their team works at the model and system level:

> So to answer that question, I have to break it down into three parts. There's the data, which is the inputs, the models, which I think you call algorithms in your language, and then the orchestration of everything, which is the system. - (P8:L27)

> …So if we keep that breakdown in mind, so with the data, you know, that you have to make sure that you get representative data, and preferably balanced, your data should reflect the group that you're studying…. So the high quality representative balanced data, that's the first step. And you feed that into the models. The models that we use in data science can be incredibly simple, just statistics, or they could be something a lot more complex. - (P8: L29)

> ...and then the system is how you add an element of human validation to interpret what's going on. - (P8:L31)

P1(L21), P2(L39), P4(L33) and P7(L21) detect bias by paying attention to both the data and the model output. With P7(L21) and P5(L36) using external tools to detect bias in the data sets thus avoid these biases from entering the algorithm. Similarly, P2(L39) performs different tasks to check bias in the data before feeding it to the algorithms, by doing this one can hope to find possible correlations. Other participants such as P6(L49) tries to identify bias by using "how analysis" on their data.

Regarding the algorithms, P2(L39) checks on the model output by asking the question "*how is it performing on different subgroups?",* as well as understanding accuracy from the objective of the system. P4(L33) points out that bias can be detected by understanding the problem that needs to be solved. Similarly, P7(L21) agrees by mentioning that an understanding of the inside and outside of the models needs to be clear:

> Understand the inside out of the model, how it works, but also from the outside in how it is applied, and how it is perceived and used by people in the actual world. – (P7:L21)

P1(L30) concludes that detecting bias can be a challenging task, thus one should treat each case individually as "*there's not two data sets, which have the same errors.".* P4(L33) agrees, mentioning that sometimes cases are straightforward and other times they are indirect.

> So sometimes it's, it's obvious in the sense, so either in the data set, analyze or in the features that you we can, okay, these might be bias, and then they need to correct this. And sometimes you have a little bit something that are more like maybe hidden to some extent, like location. – (P4:L33)

Table 10: Bias detection

Participant	Bias detecting	Comments from the participants'
P1	Detects bias in data and model results.	Mentions that it is a challenge to detect bias, as one should not generalize, emphasizing that it is a very individual process that depends on different factors as for example there should be checks, taking into account the data sets in mind. For example checking summaries of the results to make sure no one is singled out or overrepresented. P1 also mentions to check feature by feature level for combinations that exhibit unintended bias.
P2	Detects bias in data and model results.	Focuses on steps to check bias on the data before feeding it to the algorithms, where correlations should be found. Model output is also mentioned by asking the question "how is it performing on different subgroups?". Checks on accuracy are important as well as understanding the goal of the systems.
P3	No detecting	Does not detect bias, instead avoids using sensitive data such as religion and political views.
P4	Detects bias in data and model results.	Detects bias by understanding the problem definition and the input data, looking at data distributions and the results impacts, trying to understand what are the algorithms getting wrong by analyzing the features. Sometimes this is straightforward and sometimes it is not.
P5	Detect bias in data using external tools	If the data sets contain demographic information, algorithms to detect demographic parity in the data can be used to make sure there are no biases. Emphasis that if bias were to be found in the data it would not want the model to be biased too.
P6	Detects bias in the data	Explains that there is a use of 'how' analysis in order to identify bias where data for example can be split.

P7	Detects bias in data by using external tools and model results.	Uses technical metrics of bias in fairness. Data can be added and it can tell you if there is bias towards certain kinds of classes that you've defined in your data set. It is necessary to check the outcomes and understand if bias happens.
P8	Detect bias in model and systems level	Discuss different levels where bias can be detected, data, model and systems, once data quality is checked, it is fed into the models. Mentions that understanding the models is important, finally a human element of validation is added to the system to be able to interpret any output.

4.5 Real life cases

The following section will provide the answers that were given by the participants when their opinion was wanted for three real life cases. The cases will be found in the Appendix 1 - 7. Each case was presented, and the respondent had a chance to explain their answer, based on their answer some participants were provided with a new slide for the same case with additional information and the opportunity to change or add new things to their previous answer.

4.5.1 Case 1

All participants agree that data bias is the main cause for the passport system rejecting the picture that was provided by a man of Asian descent. However, the participants had different thoughts regarding what elements caused this. P1(L45), P3(L46), P5(L58) and P8(L74) explicitly state underrepresentation of data to be the cause while P2(L58), P4(L46), P6(L73) and P7(L41) mention it as data collection without going further into the details. P2(L58) explains that the system that is being used for image recognition is most likely not trained on data from people with Asian descent but that this could also be because of bad model evaluation. As stated P3(L46) agrees with it being the data collection but adds that it could also be caused by bad data preparation, nevertheless P3(L46) rounds up the answer by saying that it is because of the system not being exposed to data that is representative enough.

P4(L48) argues that this is caused by a combination of bad data collection, not good enough data preparation and evaluation. Only one of the participants changed their answer based on the additional information that was provided, which was P7(L41). After reading the first slide, P7(L37) said either data collection, post processing or evaluation. When the second slide containing the additional information was presented the participant changed the answer to only data collection. P8(L74) provided the most detailed answer arguing that not a diverse enough data set caused this leading to the training set representing the shortcoming and therefore the model evaluation could not be done correctly resulting in the data set turning the model biased.

Table 11: summary of the answers for case 1

Participant	Cause of bias	Comments from the participants'
P1	Data collection	Due to underrepresentation
P2	Data collection & some model evaluation	Data collection. The image recognition system is not trained with data from people with Asian descent, could also be due to model evaluation but mostly data collection
P3	Data collection & data preparation	Considers data collecting and data preparation. The algorithm cannot properly detect the image because of the lack of representative data
P4	Data collection, preparation and evaluation	A combination of data collection, preparation and evaluation.
P5	Data collection	Data collection. Due to underrepresentation of Asian people in the data set. The labeled pictures it's been trained on that weren't representing the facial features of Asian people
P6	Data collection	Data collection
P7	Data collection	Data collection or post processing or evaluation. After the additional fact the answer is changed to only data collection
P8	Data collection & model evaluation	Lack of representation of people with Asian face features in the data. Adds that model evaluation also is part of it, saying that the training set was not diverse enough. Their model probably did not evaluate the accuracy of the model correctly since it was not diverse enough. The data set turned the model biased.

4.5.2 Case 2

For this case the answers provided by the participants varied a bit. No participant said that it was caused only by algorithmic bias, however four out of eight participants identified data bias as the main reason for bias (P1:L51; P2:L62; P3:L50 & P5:L62). P1(L51) did however change the answer after being presented with the additional pieces of information, changing the answer from algorithmic bias to data bias. The remaining 4 participants (P4:L50; P6:L80; P7:L45 & P8:L90) agreed that it was a combination of data and algorithmic bias. P2(L62) did answer data bias but expressed the desire to add data generation as a subsection to data bias. P3(L50) and P4(L50) agreed on data preparation being the main issue in this case and P4(L50) continued by stating that skin color should not be an attribute in the data set. Just like

P1(L53), P4 changed the answer when presented with the additional information, leading the participant to add model evaluation to the answer.

P5(L62) also answered the question with data bias arguing that this was probably due to historical bias in the data indicating that dark-skinned people committed crimes more. P6(L80) also believes historical bias to be one of the reasons for the system to act this way leading to data collection being one of the main issues, questioning why attributes regarding age and criminal history are included. In addition to this, the participant believes that model development is at blame. However, the participant changes the final answer to data preparation and model evaluation, arguing that this could have been detected and avoided in the model evaluation (P6:L92).

In addition to algorithmic bias, P7(L45) brought up a discussion regarding fairness metrics saying that:

> … the problem here in this case was the fairness metric that was used … there are 21 different different kinds of ways to … mathematically … define fairness. And many of them or not, most of them are mutually exclusive. So in this case, one fairness metric was used. And the people that analyzed this case said that it was biased in another fairness metric. But it is impossible to have it balanced in both fairness metrics. So they had to make a choice. And that choice was made in the model development that this fairness metric they are going to use. And so it.. result, the result was on one type of fairness metric that it was dark skinned people were more often classified. But on another it was equal to white people. And, and that was the difficulty in this in this case. - P7(L45)

Just like P7(L45), P8(L86) was also familiar with this case and had followed the trial. The participant stated that the reasons for this were the data and model development and continued by arguing that even the data scientists that were brought in to explain the flow of the system could not do that and that the data did not have a fair representation of dark-skinned people. Emphasizing that:

> … by lack of data, representing black people with good behavior, you know, because statistically, they collect data from disadvantaged neighborhoods that have a history of bad behavior, because no one wants to be poor, but poverty forces you to do certain things. So it just reinforces the things like the wealth gap, societal gap, status gap. - P8(L90)

Table 12: summary of the answers for case 2

Participant	Cause of bias	Comments from the participants
P1	Data bias	Participant did not pinpoint a specific stage in the process, however their answer was changed from algorithmic bias to data bias,
P2	Data collection	Data bias, the participant wants to add data generation as a cause of bias as well. You could have perfect data but still have bias

		since the bias is reflected in our society and the data that has been collected will reflect that as well.
P3	Data collection and preparation	Data collection but mostly data preparation
P4	Data preparation & model evaluation	Data preparation, the skin color should not be an attribute. Could be wrongful use of the attributes. Participant adds model evaluation after knowing the additional fact
P5	Data collection	Identifying historical data as the cause, the data may indicate that dark-skinned people have committed more crimes.
P6	Data collection, preparation & model evaluation	Argues that data collection is an obvious issue due to the historical bias. A bit uncertain of data preparation but says that model development is definitely an issue in this case as well. The participant questions why attributes such as age and criminal history are relevant. Changes answer to data preparation and model evaluation because it could have been checked in the stage of model evaluation but the data preparation also plays an important role (L92). Also highlights data collection as important (L90).
P7	Model development	Model development due to fairness metrics - skipped the additional fact due to time limitations and the person having knowledge of the case
P8	Data collection & model development	The participant was familiar with this case and followed the trial. States that data and model development were the reasons. Adds that in the trial data scientists were brought in to explain the flow but they could not do that. Lack of data representing dark-skinned people was also an issue Emphasis on it being due to lack of data regarding dark-skinned people with good behaviour because its more common to collect data from disadvantaged neighbourhoods where bad behaviour is more common (L90).

4.5.3 Case 3

For this case the opinions differed more than the previous two cases. P2(L68) was the only one who answered data collection, however the participant later stated:

> … historic bias in who can be interested in technical things and who can study this? … there is probably a data that women are less likely to open some of these ads. But that does not mean that we want them to be shown less. So it's again, it's kind of like data data generation, you could almost say. But but you

> could probably have the same thought I had there about maybe that's what we should call algorithm bias and that you have not kind of taken that into account when you try to develop your model. - (P2:L68)

P3(L54) said data bias when presented with the first slide but changed the answer to both data and algorithmic bias when the additional piece of information was provided. P5(L66) and P6 stated both data and algorithms to be the reason. P5(L66) and P6(L101) shared similar reasoning arguing that the reasons are data collection and preparation as well as model evaluation since this was an opportunity to check it before the model deployment.

> … men had STEM careers in a higher percentage than women. So then the algorithm thinks that the men are more likely to click on it. So it's an issue of the data they were collecting. But this could apply to the other examples to data preparation, because they could have preprocessed this and noticed this, and then fixed it before doing the model. – (P5:L66).

When the additional piece of information was added, P6(L102) added model deployment and stated that this case sounded like algorithmic bias.

Later the topic of unsupervised learning was brought up by P7(L51) stating that this system is using that technique to maximize click rates leading to its focus on men because of the content. The participant continues by saying that it depends on what kind of ad it is but that the model is essentially striving to optimize engagement.

Lastly, P8(L112) did not give a clear answer and had some trouble understanding why the additional information was relevant. The participants assumed that the click rate is essential for making profit and that they therefore would target a cheap audience and they are expected to click more frequently (P8:L112).

Table 13: Summary of the answers for case 3

Participant	Cause of bias	Comments from the participants'
P1	N/A	N/A
P2	Data collection & algorithm bias	Says it is probably historical bias. It is a tricky one, the participant argues that it could also be called algorithmic bias. Participant did not specify where the bias came from.
P3	Model development	First answer is data bias but changes opinion when presented with the additional information adding algorithmic bias to the answer. Mentioning algorithms are being manipulated for a desired outcome which can affect the processing thus the outcome itself becomes biased.
P4	Model development	Is algorithmic bias, however, there is not a pinpointed stage in the process. Mentioning that in some cases depending on how the model is built there is not much that can be done in the data bias. Some problems are more about business value and the definition of the model objective.

P5	Model development	Data collection & data preparation & model evaluation - Changing the answer to model development, since the model was actually built to work in that way.
P6	Data preparation, model evaluation & model deployment	Data preparation and model evaluation. Motivates the choice by saying that in this case one can detect it during model evaluation but data preparation can be the reason. When the additional information was presented the participant added model deployment to the answer saying that it sounds like algorithmic bias.
P7	Model deployment	Model deployment. Mentions that this system should be using some type of unsupervised learning. It strives to maximize clicks so when certain content the likelihood of men clicking is higher, therefore the model focuses on men. Says that it depends on the ad. Reflects on questions such as what ad it is, what is the content? Says that the model is probably built to optimize engagement.
P8	N/A	Assumes that click rate is of highest importance to make profit, they target a cheap audience, and they are expected to click more often and lead to more money.

4.6 Summary of findings

The below tables show a summary overview of the participants' comments on the data and algorithmic effects on the bias for AI systems. A more detailed overview of each specific participant's comments can be found in Appendix 18 for data bias and Appendix 19 for algorithmic bias.

4.6.1 Data bias

Data collection and data preparation were highly mentioned by the participants as the key contributors of bias in AI systems. In the table below an overview of the key points regarding the data stage and its effect on bias that were brought up during the interviews is provided.

Table 14 Overview of data bias from participants perspective

Data bias	
Data collection	**Data preparation**
• It's often the data that is to blame for bias. Open-source data sets often contain bias. • It is mentioned to be important to have an infrastructure that supports companies with data generation for relevant data.	• Data scientists don't have control over the quality of the data that is given to them, as they are rarely part of the collection stage. Therefore, their work for ensuring the data is of high quality is important.

• In some cases the data is produced by the customer, therefore the power of controlling the quality of collection can be limited. Data can also be provided from a client or external company. • Important to understand the problem one is trying to solve to know what data should be gathered. • Additionally, it is important that the generated data is relevant and representative enough for the case and the place it is going to be used in. • It's good to be observant when collecting data and not collect more data than necessary. • Some companies are interested in low quality data, because their systems are used for detecting bias.	• A measure to be taken is to check the data set to see if anyone is singled out, overrepresented, underrepresented or finding proxy variables and trying to find correlations. Graphs can be helpful to understand the data distribution and finding outliers. Preprocessing is also helpful to make sure the data is good and not missing any important data points. If there are missing data points it's important to have a process for how to deal with it. Sanity checks can also be used which is hard to do sometimes since it can be arbitrary. It is important to have people with knowledge in the domain cooperating with the people working with data when doing sanity checks. Additionally, data labeling is important and a risky stage where a lot of bias could creep in. • There is an emphasis on making sure the data distribution matches the real world and is representative of the users. Lack of representative data is a common problem. To be able to do this it is important to understand the problem one is trying to solve. Anomaly detection is also important to address and can be used to detect problems regarding imbalanced data • High quality data is defined differently depending on the case.

4.6.2 Algorithmic Bias

Model development was mentioned by all participants as one of the main reasons for algorithmic bias. Followed by model evaluation which can be clearly linked with the data that is being added to the systems. Model post processing was not mentioned by any of the participants. Finally model deployment was less talked amongst the participants. In the table below an overview of the key points regarding the algorithm stage and its effect on bias that were brought up during the interviews is provided.

Table 15: Overview of algorithmic bias from participants perspective

Algorithmic Bias			
Model development	**Model Evaluation**	**Model post-processing**	**Model Deployment**
• Bias introduced by the data could be combated during the development phase. • Models are defined, built and manipulated on a specific way for a desired output, which could affect the processing and	• Is strongly tied with the data used to train, thus bias introduced in the data can be spotted in this stage, if nothing done to avoid this bias, model evaluation then becomes part of the problem. • Checks can be done at this stage to make sure the algorithm does not discriminate. For instance,	The participants did not mention this stage of the process	• There are certain places where the deployment of AI systems can be dangerous as automatic decisions can lead to disadvantages to some parts of the population, thus

cause bias • In the model development one can either remove or add attributes and variables needed to be able to deal with bias in the process. • Fairness metrics can cause bias, for example in case 2 a fairness metric used in the system where one fairness metric favored one side rather than the other one. • Models are often incomprehensible, thus making it difficult to understand where the problems of bias lay in the system. If the creators of the models are not able to explain them, that itself can be an introduction of bias into the systems.	models should be able to be evaluated regarding their performance on subgroups. • When doing testing for example with adversarial testing, one could realize that the model that something was wrong • Checks here can be used to measure the quality of the steps in the data preparation and data collection. • In this stage one can identify if the model is reading all parts needed or choosing to read specific parts. All label information needs to be used. • Accuracy evaluation can lack diversity thus contributing to bias already found in the training sets. Thus, accuracy should be tested using different classes or variables to make sure it performs wells in important features. • However, accuracy can be highly linked with profit if a model is accurate does not mean that is fair.		systems can act unfairly. • When a same model is deploy in different locations to the ones they were originally intended to can lead to bias leaving target populations more susceptible than others • It is important to understand where a model is used, referring to using a model in an unpredictable and changing environment might lead to different outcomes • Model deployment cannot be just how the algorithm is programmed but also by how it looks and how is tailored to certain demographics when deployed

5 Discussion

The following chapter discusses the empirical findings together with literature. This chapter is divided into two subchapters focusing on The risks of bias in AI systems and Data as a key contributor to bias in AI systems.

Despite mostly agreeing with the definition presented for AI (Mikalef & Gupta, 2021), participants had their own reflections and takes on how AI should be defined. This is also confirmed by the analyzed literature stating that there is no universally accepted definition of the concept (Russell & Norvig, 2016). One can however say that their starting point was the same, finding and learning different patterns from data. Interestingly enough none of the participants, except for P5 reflected regarding AIs ability to act humanly and achieve the same level of intelligence as humans. However, in the literature it is often highlighted that it should be imitating human intelligence (Russell & Norvig, 2016; Haenlein & Kaplan, 2019). Oddly enough, imitating all human behaviors might not be a desirable outcome for AI systems as these systems keep amplifying human biases thus discriminating minorities, which is normally discussed in relation to AI's dark side (Mikalef et al., 2022**;** Grewal et al., 2011; Zeng & Wu, 2021; Cheng et al., 2022). Mikalef et al.,(2022**)** and Grewal et al., (2011) point out that there is a need to address the dark side of AI, as the benefits of AI are often being talked about, overshadowing the risk associated with these systems. All participants consider it to be important to be aware of the dark side of AI as systems acting biased can have terrible consequences on people's lives.

However, even though people are aware of the risks that comes with implementing this technology, not all participants consider this when doing their jobs. Regulations are forcing the companies to act in a certain way, where the dark side of AI is being addressed merely to comply with laws in order to avoid fines (P1& P5). It is important to understand the key contributor of bias and where the risks are more likely to appear to be able to properly find, address and mitigate these biases on the systems.

5.1 The risks of bias in AI systems

This research has identified that bias can be found in different stages of different kinds of AI systems (e.g., Angwin, Larson & Kirchner, 2016; Balayn, Lofi & Houben, 2021; Griffiths, 2016; Dastin, 2018; Nasiripour & Farrell, 2021; Saxena et al., 2018; Obermeyer et al., 2019)**.** Participants have highlighted that something as small as grocery shopping can be at risk for bias. This example might not be as discriminating as the ones showcased in the literature review. However, it does highlight the importance of identifying the risks of bias in all systems. For instance, if a simple system used for predicting customer behavior for a grocery shop can trigger bias, imagine what consequences bias can lead to when found in AI systems producing decisions that could be harmful to people.

Following the sources of bias and different stages where biases can be found discussed in both the literature (Suresh & Guttag, 2021) and with the participants, we have analyzed the risks of the systems by breaking it into pieces, identifying where the greatest risks of bias occurs.

The risk associated with data collection is highlighted by one of the participants as one of the hardest places to identify bias in. The literature covers historical and representation bias, as

main sources of bias in the collection of the data (Ferrer et al., 2021; Hellström, Dignum & Bensch, 2020; Suresh & Guttag 2019). Agreeing with the literature participants selected historical and representation bias as the key sources of bias in not just the data but in the whole system. The risk for this stage can be aggravated due to companies' lack of control on how the data is being collected, in some cases using open-source data sets or relying on clients' poor quality data. Additionally, both literature and participants agreed that data is often collected where not all populations are represented (Olteanu et al., 2019). These kinds of biases can be reduced as data is being processed in the data preparation stage, however this stage is where most of the biases were pointed out by the participants. Since it is the company's responsibility to make sure the data they are using is of high quality, failing to do this can lead to bias being identified in the data preparation stage. Data preparation not only can contain historical and representation bias, but it can also add measurement bias originated in processes such as data labeling (Baker & Hawn, 2021). The latter being pointed out by participants as a risky stage where bias can easily slide into the system. Sanity checks should be performed to be able to mitigate these biases in the data. It is however a hard task to make sure to both match real world data but also be representative of all users. Thus, we have highlighted both data collection and data preparation to be stages of high-risk for bias.

Aligned with Holstein at al., (2019) views, participants highlighted that their main concern was the data affecting the algorithms. Nevertheless, even though data is a great contributor to the bias found in AI systems, it is also important to look at the algorithm whole process, as it is about tying everything together. Many of the risks directly related to algorithms were found at the model development thus we have chosen to place it as a high-risk stage**.** Bias in the model development can be found from a variety of sources such as aggregation bias which points out the problem of “one size fits all”, where the necessary variables should be taken into account when building the models (Suresh & Guttag, 2021), as for example P2 mentions when measuring strength for grades in gym class, if gender is not taken into account females would perform badly, Additionally, the literature talks about learning bias occurring during the model development as the way the model objective is defined and built can directly affect how the model process the data thus causing bias. Similarly choosing some metrics over others can prioritize groups thus leaning towards discrimination as P7 highlighted on fairness metrics used in the model development to favour light skin convicts. Another point mentioned is that when biases are known to be part of the data, they should be considered when developing models otherwise bias might be amplified by the systems.

For the model evaluation, we have chosen to assign it a medium risk as we concluded that if bias were to be avoided in the data and model development, the model evaluation is less likely to be a problem. Bias can be detected in the model evaluation by performing proper testing if the previous steps fail to identify these biases. It was confirmed by the literature and the participants that model evaluation is directly affected by bias found in the data (Suresh and Guttag, 2019). Participants remark that evaluation bias can be identified if the test performance in subgroups is being evaluated, however when testing data represents the same data that has introduced bias into the system little can be known about the systems performance on other groups. Similarly, if accuracy in the model is evaluated with no diversity in the data used for testing it can lead to the system acting discriminatively, as pointed out by participants, a high level of accuracy does not mean that the system is fair.

According to Suresh & Guttag (2021), there is no specific bias that can affect the model post processing as this step is used to modify the outcome already trained in a model, with the example of the system delivering 1 and 0 as answers, there can instead be postprocessing done,

where these can be transformed to more verbal answers such as approved or denied. However, one could argue that if these verbal answers are being identified from the developers' own prejudices, the outputs can be subject to bias. With this being said, no participant mentioned the postprocessing step, thus our conclusion for this step was found to be of low risk.

Finally, the model deployment was rated as medium risk of bias as models can be used for purposes that they were not intended to be used before (Baker & Hawn 2021; Suresh & Guttag, 2021). Similarly, systems deployed in different locations can cause deployment bias, as these locations are not as predictive as the ones the systems were intended to act first. For instance, some participants pointed out that fraud detection mechanics can generate discriminative outcomes in countries at war, letting them be more susceptible to being tagged by the systems than others. Furthermore, model deployment bias can also be caused by the interface where the algorithm is being deployed, which is mentioned by P7 referring to case 3, where people can be more likely to click depending on the design of the system, in that case, the advertisement.

Through our empirical research we observed there is a scale of how the risk of bias is viewed, that scale can be interpreted as low, high, or medium. These observations were interpreted through what participants were expressing, understanding from the conversations thus leading to the categorization chosen as presented in table 16.

Table 16: Risk of bias in AI systems

Bias contributor	Stages	Risk of bias		
		Low	Medium	High
Data bias	Data collection			X
	Data preparation			X
Algorithmic Bias	Model Development			X
	Model Evaluation		X	
	Model post processing	X		
	Model Deployment		X	

The problem of bias is a bigger issue that lies not only on the data versus the algorithmic bias, but is reflected in our society, making it a societal concern, as usually models are not being developed with malicious intent. Agreeing with Baker and Hawn (2021), it is unrealistic to expect the models to work on everyone, however when done well and with a good diversity in both the data and the team, one is able to work around biases and deliver an inclusive model. Hellström, Dignum and Bensch (2020), state that humans working with AI systems have the power to influence it which was also confirmed by the participants stating that diversity in the workplace could help detect and mitigate biases that are originated from humans. A diverse

team with not just different demographics but also upbringings can provide valuable insights and perspectives. These new perspectives can be helpful in identifying risk in the systems.

Additionally, it is important to mention that models should be interpretable, explainable, and easy to understand in case any problems might arise, thus, finding the source of errors can be easier. Regardless of the data being the culprit of most of the biases introduced to AI systems, it has been highlighted by participants as well as the literature that it is necessary to analyse the entire process of the systems in order to identify and mitigate the reason for the bias. One must be able to identify errors from either data or algorithmic processes. However, the process of identifying bias can be expensive and time consuming. Nevertheless, one must be able to check on every stage of the process and make sure that no bias in or out of the systems are found. One way of checking these biases is by applying human validations and checks through all stages in the process.

Finally, it is important to note that bias can be found in all kinds of systems, thus ruling out bias completely can be an impossible task. Therefore, one could argue that instead of focusing on data, models and systems that are free from bias the focus should be shifted towards trying to not exceed a certain percentage of bias. Participants mentions that depending on the objective, a degree of bias can be accepted. Similarly, the findings suggest that rather than focusing on systems completely free from bias one should aim for using data and creating models where the risk of bias is minimized.

5.2 Data as a key contributor to bias in AI systems

As established earlier, data is essential for AI systems to work and what they know is what has been taught to them by data (Fernandez et al., 2018). Unfortunately, it is no surprise that some people are favored over others by the outcomes of AI systems. Our participants confirm what researchers (Suresh & Guttag, 2021) say regarding representation by addressing how females or different minority groups in many cases are underrepresented in the data leading the AI systems to perform erroneously and often in a discriminatory manner for these groups, thus leading the AI systems to present us with unfair decisions. Often it is hard to control unintended bias and despite trying to control it, it has a way of sneaking in. As it has been evident in the literature (Fernandez et al., 2018; Frawley, Shapiro, & Matheus, 1992) collecting data is the initial stage of the process when developing AI systems. This means that biased data will affect the entire process, risking the outcomes of the systems to be biased as well, since they are based on biased data**.**

Most of our participants are in line with the findings from the literature (Barocas & Selbst, 2016), stating that historical bias and the problem of bias being embedded in our society is a common issue in regard to data bias. It is evident that there are many different elements to consider when collecting data as well as preparing it for usage by the models. When talking about biases in AI systems, there is a high risk of these biases appearing through data but can also be caused by the algorithms. However, it is evident that there are more crucial steps and more people involved in the stages to maintain bias-free high-quality data, making it a higher risk for bias as it seems to be easier to influence this part. It appears to be simpler to control bias in algorithms rather than in data sets for reasons such as bias being embedded into our society. We have identified data as a key contributor to bias in AI systems, thus ensuring bias

free or at least trying to mitigate the risks of bias occurring in data is of highest importance. To do this, it is important to make sure the data that is being used is of high quality.

The conclusion can be drawn that based on what kind of project it is, the data, the company, and the process for maintaining data quality is different. In cases where the data is handed to the companies by their clients it can be hard to influence the collection, but some provide their clients with requirements that they need to follow with the goal of receiving higher quality data. In one of our cases the matter of proxies being misleading was talked about, however the participants did not reflect a lot regarding proxies and the focus was more towards data labeling, preprocessing, and making sure the data sets are representative, which all are ways to ensure the data is of high quality. Furthermore, studies like Baker and Hawn, (2021), confirm the importance of using the right data for the right project just like our empirical findings show.

Additionally, an important finding that is worth further attention is the example of the apples showing us that some things are very hard to control in terms of data quality since a lot of different people are involved in the process. For that example, one would need to go all the way back to the supermarket and explain for both customers and the cashier that despite the different apples having the same prices, adding them in one bag would complicate things.

The literature presents a set of main dimensions to follow for assuring high data quality (Batini et al, 2009), interestingly enough none of the participants brought these up. Further, P6 had trouble defining a good data set, but gave some examples of what was relevant for their current project, the chatbot, this is similarly stated by Han, Kamber & Pei (2012), bringing up that the quality of data is dependent on its intended use. The literature (Batini et al., 2009), points out that as the systems change and grow there is a need for more and new high-quality data which was also confirmed by P6 when explaining that their chatbot is working together with a French company and the need for fine tuning the application with relevant data to properly work together with the French language and phrases as well.

The observation that Suresh and Guttag (2021) did regarding practitioners not always being able to collect their own data was confirmed as five out of eight participants received their data from outside their own organizations. Lastly, Olteanu et al., (2019) stated that attaining quality data was desirable but our findings showed that some projects actually needed data that represented our world correctly rather than high quality data free from bias.

As showcased, there are many different elements to take into account when working with the data that is intended to be used for AI systems. There are several steps that need to be performed on data and they can all affect its quality which could either lead to helpful AI systems by providing us with effective and fair outcomes or resulting in systems that produce discriminatory and harmful decisions, making life harder for parts of the population.

6 Conclusion

Bias in AI systems is a topic that is winning more and more attention, where regulations are being brought up to increase awareness and compliance. Bias in AI systems has an impact on society; thus it is necessary to increase awareness and involve researchers and practitioners to work together toward better AI systems.

This paper started with a focus on bias in AI systems, where the literature had an emphasis on addressing the dark side of AI by approaching one of the major risks of AI, which is biased decisions. Studies often refer to AI systems acting in a biased way from an algorithmic approach. However, we intended to demonstrate that the problem might not be solely focused on the algorithmic bias. The focus has been to differentiate data bias from algorithmic bias by looking at different literature as well as real-life examples where AI systems acted biased. By conducting an empirical study, we focused on understanding the views and experiences of practitioners within the field.

The goal was to understand how data affects the outcomes of the systems and to identify where in the process the bias occurs. By doing this we wished to highlight the importance of unbiased and quality data in systems that are using AI. We discovered that data is a key contributor to bias in AI systems. Additionally, we analyzed the literature and findings to be able to identify the sources of bias and the risk level of each stage in the process. Thus, answering our research question “What identifies as a key contributor to bias in AI systems, and where does the risk lay?”.

It is important to recognize that bias cannot be prevented 100% of the time, however, there are actions that can be taken to reduce the risk of bias in the systems. Additionally, it is important to note that the process of finding bias, does not only lie on the data, but the algorithms play an important role in identifying the bias that can be introduced by the data. In addition to this, it was demonstrated that a diverse team is able to identify bias in different parts of the process.

With the increase of technology acting in a biased way, this topic needs to be made more public and addressed from a regulatory perspective, as the everyday person often is unaware of the consequences of these systems. Additionally, if laws are not yet matured to address bias fully, the people working in the industry might not take the risk seriously. Researchers should be given the role of further investigating the subject to increase awareness, come up with unified protocols and definitions that can be used in the industry. We believe that being aware of the risks of bias affecting different parts of the data and algorithm is a first step toward combating bias altogether. Machines should not act in order to replicate today's society, but instead should be built towards making progress within society.

6.1 Implications for future research

The topic of bias in the context of AI systems is a fairly recent topic that needs further addressing not just from the perspective of practitioners but also from the research community. We have identified key points for practitioners and researchers to consider for future work.

6.1.1 Implications for researchers

Due to the extensive scope of the area of bias in AI systems there is an increasing opportunity for further research.

- **Dark side of AI:** It is important to increase research on the possible risks of using AI systems. The goal should be to outline not just the benefits, but also the drawbacks of these technologies to be able to encourage a debate in the field, thus leading companies and governments to act on these drawbacks.
- **Mitigation practices:** Designing clear principles on how to mitigate bias in the mentioned stages of the systems should be further developed by addressing each of these stages individually with the help of practitioners' best practices.

6.1.2 Implications for practitioners

AI systems are affecting the lives of millions of people; thus practitioners should take into account implications to avoid AI systems being discriminative.

- **Bias awareness.** Our findings point out that attitudes towards addressing bias in AI are different depending on the specific industry the participants are working in, as well as their roles and knowledge. Participants that had a deep knowledge of the risks of bias took this into consideration when working. As opposed to participants who didn't have any deeper interest in the topic. This leads us to believe that further training needs to be taken into account for companies and employees that develop and use AI systems.
- **Increase support for developers.** It was noticed that there is a need to support developers to address bias in AI systems. The topic of bias is fairly recent, with regulations appearing and evolving, the responsibility of addressing bias should not fall on the shoulders of the developers only. Bringing experts to assist the team of developers in parallel can have a positive impact on the company.

To conclude, it was noticed that joint practices by both practitioners and researchers is highly encouraged. Especially as these systems are being rapidly developed. We believe that including experts on the fields of bias in the development and evaluation of models would be very beneficial and will help decrease the risk of bias in the systems. As well as including practitioners' views on research can allow for deeper insights on this rapidly changing environment.

Appendix 1: Interview guide

Dear Participant

Thank you for taking the time to take part in our research. We appreciate your participation and we commit to utilizing the information from this interview in a responsible and secure way.

About us and our research:

We are Information Systems Master students at Lund University, currently working on our thesis. Our research focuses on Bias in the context of Artificial intelligence systems from a data perspective. The questions of this research are based on the knowledge we have obtained during the process of writing the thesis.

Your rights:

The interview should take around 1 hour. You have the right to stop or withdraw from our research at any time. Your participation will be anonymous. If you give us consent the interview will be recorded and transcribed. The transcription will be sent to you, to give you the opportunity to correct any errors or add any insights you may have. Finally, the recordings of the interview will be stored in a secure location and destroyed as soon as the transcription is processed.

Warm-up:

1) Please provide us with a quick introduction about yourself

2) Can you tell us about your experience with data and artificial intelligence?

Areas for our questions -

AI view

1) There are many different ways of defining AI, we define it as: “AI is the ability of a system to identify, interpret, make inferences, and learn from data to achieve predetermined organizational and societal goals.” Do you agree with this definition? Would you add/change/remove something?

2) The literature that we have read points out that there is a need for addressing the “dark side” of AI such as the risk of AI systems acting biased. Is this something that is talked about at your workplace and among colleagues?

Data view

1) When collecting data, do you have specific requirements that you follow or other aspects you try to consider?

2) How do you make sure that the data that you use is of high quality?

3) Can you describe the process from when you get access to the data you are working with, until it is used as input for algorithms?

4) Can you ensure there is no risk of bias in your data, algorithms or the systems you work With?
→ If yes: How do you ensure there is no risk of bias in your data/algorithms/systems?

5) Do you work in a more precautious way in relation to bias when working with data that contains demographic information?

Bias view

1) How do you work in order to detect bias?

2) Is your organization able to identify bias, and if so, what type of biases have your organization identified?

3) Based on your previous experiences, what is the most common type of bias?

Team diversity

We recognise that team diversity plays a central role in bias detection as opposed to working individually in bias detection. Therefore we want to proceed with the following question:

1) How diverse are your teams? Do you think having diverse teams can have a positive effect on detecting bias?

Real Life cases:

During the interview, you will be presented with 3 short real-life cases of AI systems acting biased. The reasons for the bias will be discussed.

Appendix 2: Case 1

Interview

Case 1

A man of asian descent wanted to apply for a new New Zealand passport online. The photo he uploaded was rejected and he was presented with an error message that said his eyes were closed and the photo did not fulfill the requirements of the system. His eyes were open and after three attempts with three different photos he had to reach out to the passport office. The office blamed it on bad lightning and shadows in the eyes.

Data Bias
- Data Collection
- Data Preparation

Algorithmic Bias
- Model Development
- Model Evaluation
- Model Postprocessing
- Model Deployment

Appendix 3: Case 1 - Additional information

Interview

Case 1 – Additional information

There was an underrepresentation of people of Asian descent in the dataset used to train the model

Data Bias
- Data Collection
- Data Preparation

Algorithmic Bias
- Model Development
- Model Evaluation
- Model Postprocessing
- Model Deployment

Appendix 4: Case 2

Interview

Case 2

An algorithm called COMPAS was used in the United States to determine whether or not a defendant that is awaiting trial will re-offend. The decision was based on more than 100 different factors such as age and criminal history. When analyzing the results it showed that dark-skinned people were more often classified as high risk among the defendants that did not commit new crimes.

Data Bias
- Data Collection
- Data Preparation

Algorithmic Bias
- Model Development
- Model Evaluation
- Model Postprocessing
- Model Deployment

Appendix 5: Case 2 - Additional information

Case 2 – Additional information

- Granularity of data was different
- Used the prior arrests status of the defendant's family and friends to determine the likelihood of them re-offending
- Not unusual for minority communities to have a high police presence, leading to a higher number of arrests. Drawing the conclusion that the residents of minority communities have a higher number of dangerous people due to having a higher number of arrests is misleading due to factors such as the police presence.

Data Bias
- Data Collection
- Data Preparation

Algorithmic Bias
- Model Development
- Model Evaluation
- Model Postprocessing
- Model Deployment

Appendix 6: Case 3

Interview

Case 3

In 2018, it became known that ads related to STEM careers were less likely to appear for women than men. The ads were run on many big different social media platforms such as Facebook and Instagram. However, women were less exposed to the ads, and they were appearing more frequently to men.

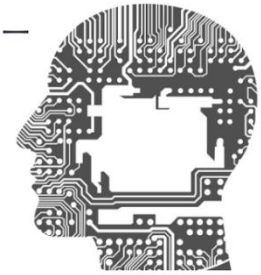
Data Bias
- Data Collection
- Data Preparation

Algorithmic Bias
- Model Development
- Model Evaluation
- Model Postprocessing
- Model Deployment

Appendix 7: Case 3 - Additional information

Interview

Case 3 – Additional information – possible reasons



- The algorithm learned the pattern from the consumers. Women were less likely to click on the ad therefore it trained itself that showing it to men would be of higher value.
- Presenting ads to women costs more and advertisers must pay more for exposure to women. The algorithm is programmed to make the most cost-efficient choice and is therefore leaving women out.

Data Bias
- Data Collection
- Data Preparation

Algorithmic Bias
- Model Development
- Model Evaluation
- Model Postprocessing
- Model Deployment

Appendix 8: Transcription - Participant 1

1.	Speaker 1	Okay. Thank you. Okay. So we have some warm up questions just to get to know each other better. So the first question is, if you could just give us a short introduction of yourself, and what you have done so far, what you work with.
2.	Partici-pant 1	Yeah, so my name is X, I work here in Risk Advisory at X . I've done that for about a year and a half. Before that, I studied mathematics and economics in X, and I did my master's thesis on reinforcement learning and on the ability of those algorithms to generalize. So my education has been very quantitative. And my so my focus has been on machine learning for the majority of my master's degree. I went and started here in risk advisory in an innovation team. That's where I started working in like in non academic sense with machine learning. So my career began yet. So at company X , we have a, it's referred to as an incubator. In Frankfurt, in Germany, called X where they are, I think, nine or 10 countries who are participating, they each send a data scientist there. And we then collaborate to build tools to assist either as either to be sold or software to assist consultants at company X. And when I was there, I worked on a tool called AI qualified, they changed the name since then **Smiles** . But which is about the robustness of machine learning algorithms. So like testing their ability to maintain the correct decisions when met with like certain amounts of noise or similar **hands movement** . That is a part of a wider framework called trust-worthy AI at company X, which includes the device part and the trans-parency part **hands movement**. So that's my way to reach them as part of machine learning. I've done a bunch of other stuff at company X, but this is the machine learning half of it.
3.	Speaker 1	Yeah, sounds exciting. Our second question is, if you could tell us about your experience with data and artificial intelligence. I feel like you kind of did that. But is there anything else you want to add? Before we move forward?
4.	Partici-pant 1	No, I think so. Yeah, my background in this very mathematical. And my contribution to these projects have been in the technical sense. I have an SS, what I was building or what we were coding was a tool to assist. We weren't actually looking at any client data or anything. We were just we were approaching this from a very theoretical standpoint for them, just how would we build something if we had that in hand, and build something that would generally be useful, regardless of what the application was? Okay. Yeah,
5.	Speaker 2	I think that is also sound super useful right now is needed **participant agrees and smiles**. After our research, we can realize, like, we see so many examples that nobody sees that, okay, this is actually added bias, and we didn't even realize that it was actually bias **participant agrees and smiles**. So, for these, like we are starting our research. So we

		have, we're continuing with the questions of artificial intelligence. So we had 2, and we define artificial intelligence as the ability of the system to identify, interpret, make inference, and learn from data to achieve determinate organizational and societal goals. So do you agree with our definition, I know it's not very mathematical, and it's more like society kind of definition but if you will have anything to add or remove Are you okay with this definition?
6.	Participant 1	I think it makes sense. It's very hard to define AI mathematically because it, it's so many things at this point, you're, like, ranging from NLP applications to computer vision and whatever. So I think think a more societal definition makes more sense anyway. So it's a hard task to define **smiles**.
7.	Speaker 1	There are several different definitions, like based on who you ask and what their field studies and what they work with. So, yeah. **participant agrees**
8.	Speaker 2	But that's great, then that we got to my second question, then. And so the literature that we have read, points out that there is like, the need of addressing the dark side of artificial intelligence, because of course, I so many people say, and has been talking about how good artificial intelligence is without actually knowing the risks **participant agrees**. And so is this something that is talk about at your workplace and among colleagues, I know well, from the trustworthy AI **participant agrees and smiles** of course, it is taught, but like, in what ways been taught.
9.	Participant 1	So since I now work in the mine, in my team we are very bit like we are addressing this from like a regulatory viewpoint. Since the regulations haven't really caught up with all the, like, the EU has made some proposed regulations stuff, but there's no like concrete, this is how and this is how you're going to get fined. As far as I know, as least. It's not an immediate like big ask this my impression. However, we're very interested in seeing because we do expect this to be a big concern in the future, when we do have like concrete numbers, and we see companies get big fines because of these topics. So the interest is, from our side is very much on both like in model validation perspective. Where I think all of these trustworthy AI topics are very, very much like, like a core part of doing model validation, apart from your standard, like performance tests, but this should all be integrated. And then we see this, when we've spoken to clients, that concern from this is often that you have your model development team, which are working very hard to build models already. And then comes all that but now you have to do bias checks, you have to transparency, whatever, you have to do robustness, testing, whatever, and they were too bogged down development teams. That's why as a consulting firm, you have like an option to contribute. Because you can do this, like this, does this can be done in like a parallel way, and not bogged down, or give tons of more work to developers. Yeah. So that's, that's how we are discussing it. We are very focused on the robustness still, but I because that is one of the most that's where I have the most experience and have I know consulting also are contributing a lot to the bias part, the bad part, as you already

		pointed out, that exists very much in the data not very often only cause exist in the data and the model is such where the robustness is more on the model side. Yeah,
10.	Speaker 1	interesting. I have a follow up question. As a person, like as a private person, but with a lot of knowledge in this field. Would you say that you personally are concerned about this, like the future of AI and it being biased?
11.	Participant 1	I'm, I mean, I am concerned. As of I would, **rethinks** I'm not very concerned because I do I am quite confident that we will have like the regulations in order to address this the course is often not like an ill intent in when you're building models. It's more like lack of attention to it. So one as this this topic is, as you guys probably know better than me it's getting a lot of interest in this this will this will not die down **smiles**. I think people will be very aware. Yeah. And once people have to pay money when they don't do it, then **laughts**
12.	Speaker 1	yeah, exactly. X, do you have any more questions that are that?
13.	Speaker 2	I think for that section, I'm fine. And maybe you want to continue with the other one.
14.	Speaker 1	Yeah. And then we're going to skip to a little more data focused questions. So I don't know if you have worked with data collection or things like that. But our first question is, if there are any specific requirements that you guys follow, or aspects you try to consider when you're collecting data for your models, I know you mentioned that you didn't really use any data for the models you worked with. But if you have an input
15.	Participant 1	We used open source datasets. So but but I think so those are just like very commonly used openly available datasets, where the the bias components are, you can Google them if you want to. It's very, it's very out there. What's in these datasets already We use hDmA data set I think home something about about home loans.
16.	Speaker 2	Okay.
17.	Participant 1	And what we did, what was the focus of that is to like, isolate out, like less representative groups looking at specific, like sensitive features there in, in this case, so this is just a dataset we use for the demo, we ran a bunch of it through it. sex, gender? nationality, were the I think, yeah, those are like the center focus of that are like what we deemed sensitive within that. Then we looked at like group from group to group within those are the demographics that are being treated less well than others? Are there any demographics that have very low, I think Eskimos or something had like two data points. Oh, and so that those type of stuff was what we looked at. In our case, since we were building tools. We were trying to make a demo, we wanted there to be stuff. We actually wanted the data set with a lot of bias, trying to demo the fact that we could show bias.

18.	Speaker 1	Yeah, okay. Interesting. So you would say that, when you want to check the quality of the data, you go through it and make those analyzes of the demographics and things like that?
19.	Participant 1	Yeah, I think you can like you that that's like your first first cause of action is to just take a very, like summary based look like, is there any one who's singled out? Is there any thing that's like, well over represented? And if it is, is it like, is it gaining a positive advantage, as such, then the next steps? Looking at, **thinks** like proxy variables. So I think, in our case, the demo shows is that, for example, etchnicity, you have something like postal code that can be quite, you have certain areas with ghettos or something where the ethnicity is, to some part reflected in the postal code of the address. And then, so even if you just like get rid of ethnicity as a feature, the bias is still inherent in the dataset. So there is a second round of checks that so those are, like, a short summary of what, yeah, what we looked at.
20.	Speaker 1	Yeah, that leads us to our next question, which you kind of touched upon, but can you describe the process from when you get access to the data? And then when you're working with it until that it's used as an input for your algorithms or models?
21.	Participant 1	Yeah, so I'm a bit on the weird side of this, because it's not really the model I was working with. I was working with this, like plugin tool **smiles**. So we didn't like we didn't do anything to it because we were actively trying to showcase a model with bias. Yeah. But as I I said like there was some So initially you there are some very quick fixes or fixes but quick takes to look at like is do you have? Like, is there only one person from Denmark? Yeah, so if that guy is having has like a bad outcome, you need to be pretty sure that that is not reflected in the model. Henceforth like, so if your model is an over fitted things, people are going to have a hard time. I mean, it's a bit more complicated than that. But that's like the basics of it. And after those initial assessments, you can like you can keep digging into this, then you once you possess this on like feature by feature level, then you can assess certain combination of features might have like a certain nationality when paired with a certain gender exhibits a kind of bias, that is unintended. And then there's the winch once you go down to look at proxy variables and stuff that's. I have, so I've limited experience with this. But I think that's like the basics of it.
22.	Speaker 1	Yeah, thank you. So our next question was, can you ensure that there is no risk for bias in your data or algorithms? But I guess you didn't really that was not the goal of your project?
23.	Participant 1	But I will say, I don't think anyone can say yes to that.yeah **smiles**, it's true. Yeah. Like you're bound to have some bias. And although like, the difference is really whether the bias is intended or unintended.,
24.	Speaker 1	yeah, and maybe humbled to the fact that there might be bias, because we have read some things about like some systems that, like, there's a high risk of them being biased. So therefore, they're used together with human decision making, just to like it could be used as extra input, but we still don't trust them enough to make the effort to let them make their own decisions **agrees**.

25.	Participant 1	Yeah, definitely. What we did as I can add to that is, so part of the analysis, the tool that I didn't program, this tool, by the way, I've just been demoing it was to instead of saying what wasn't bias, it said certain intervals of accepted bias for like, maybe you would accept bias a 5%. Better output a worse outcome better from a worse outcome. If your break within like the 5%, you deem that like an acceptable level, you could base that, or you could lower that. But setting that to zero, you're gonna have a weird model at the end.**smile**
26.	Speaker 1	Yeah. Thank you. We have one last question. Regarding the data. It's about the demographic data that you use, do you handle it in different ways from other data? Since it can contain some sensitive information? Are you like more precautious? When you deal with it, or?
27.	Participant 1	So I've only handled that **laughts**. But I can say from like, so I've worked as part of a team who audited machine learning models? Yeah, I can say I've seen clients handled things very differently where they would even have, like the law team. Yeah. As part of so when this client were building models, they would like at the end list the features they were using, they would pass that through legal and then legally would have to, to give the okay sign for these features being used. So that kind of process would like restrict, I think that they would then I don't think you will ever allowed to like use basically any sensitive features. That kind of process will make sure you don't like take the wrong steps. Yes, they will all legal team don't know everything about **smiles**
28.	Speaker 1	Yeah, but it would go through a check with them before **participants agree**. Thank you. I think that was all for the data section. You want to continue?
29.	Speaker 2	Yes, I can go more towards bias itself and actually made sense that you get to build a legal team because of course, this is going to be expensive to the systems and to the company if bias were to be found **participants agree**. And so I know you're you're basically on the other side, which is great to actually get the point of view Like, no, we're actually trying to put all these bias in the systems and kind of showcase them that they actually there **participant smiles** . And so what what do you think one of the questions is like how do you work in order to detect this bias? Like do you follow any protocols? Or would you after this model is being done. And so, will you actually say that there is a specific steps that we follow in order to realize there is bias? I know you say like, you summarize and you say there is like an a specific nationality that is, single out. Is there any, anything more like specific guidance for that?
30.	Participant 1	No, I mean, I think I really like my knowledge on that point is what I said before **thinks**. I have talked, since my experiences, mostly based on what this tool from the company does, that would be like the process I would explain **smiles**, of course, but it is very individual. That also is a big challenge when building tools like that is this is like, there's not two data sets, which has like the same errors. So, keeping that general perspective in mind gets really tough. . So that's why this is

		you should definitely have checks. Like with the data set in mind, in addition to all the the first glances.
31.	Speaker 2	it does make sense. So, maybe it goes towards a second question. So we know that your organization is able to identify bias, of course. And so we're wondering what type of bias Have you identify, like, there is like a few, of course, that we thought in our research, like historical bias, sample bias, and but there may be as well like, some bias in the model and biasing the data or Algorithm? So yeah, what type of bias? Have you identify
32.	Participant 1	**Participant thinks ** Yeah, so that sort of, again, I'm, I haven't really worked with datasets that weren't just like open source datasets, I found, we found the biases, I explained before, we had like, the Eskimo class that had one, I think, only one data point. And that data point did not get the loan or denied, then you have that type of problem. Other than that, I've looked a lot into computer vision. Where so in computer vision in general, like there's a lot of white male people, which means that both women and especially women from, like, minority, underrepresented, so you have algorithms that just function far better, if you give them like a Scandinavian guy, and sort of like drop from there **hands dropping**. Then, no, I think that's like, that's the most, then I've I think those are like, my two best examples.
33.	Speaker 2	Thank you. I think you I think that made sense. And would you then agree, I mean, not to be biased *participant smiles*. But do you agree, like thing is like, bias mostly comes from the data rather than the algorithm itself? Because, of course, if you're gonna fit bad data to assist again, gonna, the output is going to be bad, too.
34.	Participant 1	Yeah, I agree. I think it's like, if not completely, then very close to a data problem. The models are just like learning whatever you show them.
35.	Speaker 2	There, that's great, like, and then one more question towards more like team diversity. So we understand that team diversity plays a central role in detection, bias of opposite just working individually. So we just have one question in here. Like, how diverse would you say that your team is? And do you think that having diverse team can have a positive effect on the detecting bias?
36.	Participant 1	Our team is very diverse. I don't remember exactly. We have a lot of nationalities. And so I don't know the numbers for it. We are quite evenly split, like gender wise. I think it's I think it's important because it sort of gives this like subconscious understanding of things just being exposed to a lot of stuff in the workplace.So not necessarily address, not necessarily, in a direct sense, but just opens up your mind to the fact that these things exist once you get in a more varied, like diverse setting.
37.	Speaker 2	That's good. And I've been now we got to the fun part. Other interviews, so we have a few cases, like only three cases where there has been like, bias happening in real life. And so we want you to identify, of course, there is like a lot of things behind the bias **participant agrees**. So we're gonna give you a like a general view. Let me just share my screen. Hopefully, it doesn't crash us last time.

38.	Speaker 2	**sharing screen with cases**
39.	Speaker 1	It's just from your point of view. So just to give you a quick overview, so we have identify a specific part in the specific process. So we'd say like, okay, that the bias might happen from the data collection, or the data preparation, an algorithmic bias might happen, like in different states will could be model development, model evaluation, post processing, or deployment. And if you do need like a refreshment of what they are like, there is like, the, the tiny number says, like, which each one is. But otherwise, if it's very straightforward, we can continue with the cases. Okay,
40.	Speaker 1	so this is our first case, I don't know if you prefer to read it for yourself, or if you want us to read it out loud.
41.	Participant 1	I'm reading. Yeah. *smiles**
42.	Speaker 1	So just let us know when you're done.
43.	Participant 1	Yeah, yeah, hang on.
44.	Speaker 1	Yes, I don't know. Those are like real life cases. So maybe you've heard of them before. But starting from the data collection, to the model deployment of this AI system, where would you say is the highest risk of this bias occurring?
45.	Participant 1	So without knowing like, what the actual light settings were, that wouldn't be light was bad *jokes*. But I would suspect that that's, like, sort of like the case I pointed out before, like a lot of data. Picture datasets are full of white male people. And seems like certain ethnicities or certain ages even have like, kind of different facial features. This model could very likely be overfitting to just detect whether it was looking at a Caucasian male.
46.	Speaker 1	So if we scroll down to the next slide, we have one additional point of information. And I am my question would be that, is this the same? Would you answer it the same way as you did before? Or would you like to change your answer now that you have access to this information?
47.	Participant 1	No. So like I said, I think you canz, you can combat this during development. Even if you have a representative stuff you can, like you can add regularization to these photos, you can add more. Like even if you are not able to go out and photograph a bunch of people. You can still like to, like you can blow up the dataset. And in that way, so you can just synthesize more Asian people based on the stuff you already have. So it's it's in part a fault of the fact that the dataset was under I was under represented Asian people wise, but this could have been like, this is not the fault of the guy was the camera. *laughts*
48.	Speaker 1	So let's continue to case two. So if you let us know when you're done reading it **reads** So it's the same question from the data collection to model deployment, where would you say is the highest risk of bias occurring for this case?

49.	Participant 1	So I think, again, I would assume that the data is at hand here. Whether it's because there's not enough women or because they aren't like in the dataset at all, or if it's because the woman who are in the data set has been rated lowly. If it's the second case, I guess there's a problem in the whole human resources process. If it's the first case, it's like under representation and then there's a balance between those two. Which might be the issue.
50.	speaker 1	Yeah. Okay. And then we have an additional point of information for this one as well. Yeah, so I don't know. I guess you're going with the same answer as before. **participant agrees** So then we have one more case. This is the last one No, it was one.
51.	Participant 1	Yeah. Yeah, I've seen this one. Before. I actually worked with this as well **smiles**. Like we use some of this. Some of our slides. This was like clearly biased algorithm that seem to take mostly skin color into consideration, not like I've seen some pretty bad examples, not saying anything. Bad about people with face tattoos, but like, I've seen pictures of white guys with face tattoos and which, I mean, that might also be an indicator, but it clearly didn't pick up at that at all.
52.	Speaker 2	Yeah, will you l just say that was a bias in algorithmic?
53.	Participant 1	Yes yes **thinks about it** No, I would still say there was a data. I think it's, it's a data part.
54.	Speaker 1	Uh huh. Okay, and now we have some additional information for this case as well. It's a bit more than the other cases. Though. We felt there was important. **thinks**
55.	Participant 1	Yeah, I think this slide on, is a dangerous place to start using I'd like to start using machine learning in the first place.
56.	Speaker 1	Yes.
57.	Participant 1	And so as this slide clearly shows, there's a bunch of factors that that plays into this, of course, know how they collected it. It was probably a bunch of different precincts. That's how I understand that the granularity of the data is different. That they collected, yeah, you didn't have the same amount of information for everyone. You know how you treat that. Like that can that's a big, pretty big red flag. If you then like if you decide to just go forgo that data or you're deciding to like, do some other types of filling in voids. And then like the whole police presence factor is, of course, like we that's very hard to pull out of the dataset. Yeah, and it's like these are difficult topics to address model wise.
58.	Speaker 1	Yeah. And you mentioned something about proxies before as well because I know that one of the comments they made about this case was that the proxies they use were not fitting for this type of case. And that could be one factor as well.
59.	Participant 1	Yeah. Good. Yeah, definitely. Especially if you have like very granular data for some people. Those that you end up having like a one to one relationship, because you have like enough of granularity in a small enough sample size. Then your model might just correlate those things completely. Yeah.

60.	Speaker 2	I think we are done actually, with the interview. It took way. It was way faster than we thought it would be **smiles**. Which is like great for everybody but at the same time, like do you have any comments and maybe something that we can improve? Or maybe something that you think like, okay, it might not be relevant.
61.	Participant 1	No, I think I think I think this was I think the reason this went fast is also because as I like, we mentioned a couple of times, I don't have, like, I haven't had my hands on a lot of actual, like data. So we sort of skipped a couple of questions there. I think it's very good topic you guys are working with and I think the it seems like you guys have a good understanding of were the issues like that like, there are ways to like, address these issues. We didn't like we didn't talk a lot about head like handling it once you've detected it. Yeah, I don't like this. I guess your scope should also be limited. In some some ways. Yeah. I guess that is that is something if you want to have more, you could. Yeah, I think we have things might get technical.
62.	speaker 1	Yeah, we've gone back and forth with that to make sure it's in our scope. We've used that word mitigate the risks a lot, but then we have like, come back to that. Oh, no, that's probably outside of scope. But that's a really interesting thing to talk about. Because yeah, well we know all of this. So what are we going to do with it?
63.	Participant 1	I from what I've heard from a lot of very intelligent people who work on this topic is that this is this is very difficult. It's not like you can just name like free techniques and do those. It's very individual, and it should be addressed on like a problem by problem.
64.	Speaker 2	That's actually a great point to take on. And that is true, because the thing in our research is like we have, we started the research there was the data bias because the most of the articles we read was like always talking about algorithmic algorithmic bias, and it was blaming the algorithm even though they were talking about data, so we can then realize that maybe it's better if we separate both and then see, okay, how data actually affects the rest of the process. So so that's why like, we then got this final result, how we're actually focusing
65.	Participant 1	Yeah, I think that's a that's the correct angle. The models are not inherently biased. They are like just mathematical techniques to model data. So if the bias has to come from somewhere, in my opinion, it's on the data side or like a lack of attention during your model process. Because then bound data is bound to be biased to some extent. During the modeling process, that's where you would then see, like, with the knowledge that you have this bias, what do we then do?
66.	speaker 1	I agree with that. I don't know if you don't have anything to add, X?
67.	Participant 1	you have anything?
68.	Speaker 2	No, I'm fine. Come in. It's everything. Okay, from the recording point of view.
69.	Speaker 1	Yes, everything should be fine.So thanks a lot for your time

Appendix 9: Transcription - Participant 2

1.	speaker 1	Yes. And just information, you're allowed to skip whatever question you want to skip, or if you want to cancel the interview. That's okay, too. Just information about that.
2.	Speaker 2	So I'm giving you, I'm just gonna give you like a quick overview of what is our thesis. So the title of our thesis is bias in the context of artificial intelligence systems. And the focus will be to tackle these from a data point of view. So we're interested to find out how the data affects the outcomes, and of the actual day AI systems. And we want to identify where in the process, the bias may occurs. So it could be like maybe the data itself, or is the algorithm I know it might be? It's tricky to find out, but we're just trying to like, figure it out. Yeah.
3.	speaker 1	Yeah. So are you do you want to add something, speaker 2? No, I think I don't look. Yeah. And I just want to add that we will transcribe this and then we will send it out to you. So you get the opportunity to go through it and make changes if you want to. But we can start with some warm up questions. Could you give us a quick introduction of yourself, and what you have been doing and what you work with?
4.	Participant 2	Absolutely. So I have my background, actually, primarily in political science. And before I started studying, at the Information Systems program, I did a Bachelor in political science, and then kind of in parallel and a little bit, afterwards, but mostly in parallel with the Information Systems bachelor, I also studied quite a lot of statistics. And so since like, a little bit more than a year, I've been working as a data scientist, first at this grocery company in Sweden, and then now I'm working in Gothenburg at a startup, who is working with the training data used for autonomous vehicles. But my role is very much data. Related, you could say,
5.	speaker 1	Yeah, because the next question was like, can you tell us something about your experience working with data and artificial intelligence?
6.	Participant 2	And yeah, so what, what should you say like in company X, I worked in a team that focused on product recommendations and personalization. But what I did was like customer segmentation and customer profiling, you could say, so as an example, I was working with estimating the probability that an account had children living at home based on what they are purchasing. And so then you are kind of trying to build a statistical model to estimate that probability. And in some senses, that's, that is maybe a mix of very traditional statistical modeling and more of a machine learning approach to it, because you're quite focused on getting good predictions. And that's much closer related to the machine learning field than the traditional statistics field. And here, company Z I've not worked so long. But it's still very much related to instead, producing good quality training data for AI systems, you could say, and actually validation as well. It's not only used for training, it's also used to validate other models like neural networks and other AI models, basically.
7.	speaker 1	Interesting.

8.	Speaker 2	Yeah, it is. And that's the thing of course, trying to figure out if there is kids in the house, depending on where you buy, totally made sense. Yeah, like I'm not gonna buy diapers if I don't have a child at home. i I hope Yeah. Maybe we can continue with the AI view is that okay? **participant agrees** . So, we have only two questions in the AI topic, which is like one there, we realize that there is many definitions of AI and we found on the Phoenicia more social. So, we want to like give it to you and then you let us know if this is you agree or do you would like to change something or remove something? So the definition of AI and AI is that AI is the ability of assistant to identify interpret, make inference and learn from data to achieve or terminate organizational and societal goals. So do you agree with this definition?
9.	Participant 2	Yeah, I mean, I think it's really hard to pinpoint a definition for it. And I think an interesting thing about the definitions is that it only feels like artificial intelligence until you have not made it work **smiles**. Or like when maybe back in the days, it felt like super artificial intelligence to even automate something on a computer. That felt intelligent to us, because it could not yet be a shield. And now we expect kind of a lot from this system since then. And maybe we don't even think I've read this elsewhere then. But that's that it has to do with this unachievable things well, but overall, it seems correct. Like finding patterns in data, I think is a very good overall, like, explanation for it as well, that captures a lot of what you do, I would see this as an umbrella term as well. That covers quite many different fields, like robotics, and information systems and statistics. And yeah,
10.	speaker 1	yeah, if someone were to ask you what AI is, how would you explain it?
11.	Participant 2	Yeah, I don't know. I personally don't like so you use the term AI so much, or try to refrain from it, but it's quite hard, because sometimes you will just get in the loop with what everyone else is. But I prefer to talk only specifically about machine learning instead. And then maybe that's, but maybe I don't have a good answer, either. But I think that you have captured many important things there in in your definition.
12.	Speaker 2	Perfect. Yeah, machine learning is a little bit more specific And at the beginning, we were we were confused. Okay, machine learning, deep learning AI everything together umbrella. So we we can realize, Okay, let's try to use that bigger term to see how do we encompass everything together. **participant agress and smiles** So that goes to my second question. So the literary literature that we have read points out there is a need for addressing the dark side of artificial intelligence. And like the rise of AI system Atem bias. So we were wondering, is this something that is talk about in your workplace or among your colleagues?

13.	Participant 2	Yeah, where I work now we speak about that all the time, because our mission is very much related to that, like recognizing that there are so many things that can go wrong in autonomous vehicles, and that there are still very many hard scenarios that are, we don't really know exactly how, how to build good self driving cars. But that is, that's a huge impact and all of that, and then kind of really importance for therefore validating these systems. So that you can regulate, like, what kind of safety do we require from an autonomous system to actually go out and drive so to say, so for us, these dark sides are very present, but maybe at my previous job, it was not discussed as much there it was more about levering leveraging that technology. So you could say, Yeah, but it was present as well. But here in the I mean, now I work at a company that is very much focused around solving these, these tricky, these like dark side problems, so to say,
14.	speaker 1	Yeah, I guess like the level of the risks are different also, based on what you use it. Like for a vehicle it can be quite dangerous but predicting that someone should need diapers, but even though they don't is not that much of a risk. **participant agrees**.
15.	speaker 1	If we move on to the Data section, we have some data related questions. So when you collect data, do you have any specific requirements that you follow any protocols or how do you think when you decide what data to work with? What data to collect?
16.	Participant 2	Yeah, so maybe here it's good to say that where I work now, we are not actually collecting any data, but we are part of the this whole chain and ecosystem of machine learning in the sense that other companies collect data, and they come with the data to us. And we help them with labeling that data. And they could also come with the output of their system to compare with ground truth, annotate the data. So so then we don't do a lot of data collection ourselves in that sense. But of course, we have internal things that we collect data on, like our users behave and, and that type of stuff. I'm not, I'm not so familiar with the processes for that. But maybe, instead, if I roll back and think about my previous job, then then that could be many different things. Like sometimes we would have the data already, it was core part of the data we have about customers. So like, you can think about, your membership card, so you know, what the membership customers actually buy, and when they buy, and what stores and so on. But then for this thing with children, instead, we had to collect that data through a survey sent to the customers. And so sometimes, I mean, maybe you start sometimes you will have like infrastructure in the company that supports the generation of certain data. And means that you don't have other data like this the information I found households or children or not, that's like, actually, the government has that information. But as a company would, kinda didn't want to buy that information. Because we know customers don't like that either. They don't want companies just buying more information about them. So yeah, yeah.
17.	speaker 1	And that leads me to my second question, I like the data that you work with, how do you ensure that it is of high quality? Do you do any specific tests? Or how do you think about it to make sure that the quality of the data is good?

18.		This is actually a field that is like much harder than you or could many people think, I think, like a lot of companies, they just want to do AI because it's cool, but then they are not really thought through the quality of their data. And oftentimes, it's that responsibility is kind of far away from the data scientist, like as a data scientist, you just kind of receive the data. And then maybe after a while, you notice that like, Oh, this is crap, or this is wrong, or you know, and you might not, I mean, I think you can work with processes in one way, like a typical example in the retail field would be that. You know, sometimes, if you are buying different apples, you might just notice that when they have the same price, I can put them in the same bag. Right? Yeah, that will give you low quality data on the both apples because it will think that someone only bought, you know, Pink Lady apple, but actually had a collection of different. And then there's a challenging getting the store people to understand like, that's actually not the store because sometimes it's what the person in the cashier **machine sounds** into the system. But sometimes it will be like, just you know, how, how do you make it so that your customers won't behave in a way that creates bad data. That's, that's actually hard. And then you can have some other things that's more maybe like technical. So yeah, having checks in your data pipelines, and really making sure that you have assertions here or there in your code or in your database system in general, and making sure that these fields cannot be this number, or you can have more like sanity checks. And you can have more like exact calculations. And sanity checks are really hard to set sometimes, because it can be quite arbitrary.*mmm* Sometimes you will think that okay, but of course, no one is shopping more than 60 times in a month, right? That was not reasonable. But But actually, if you're two people in your household, and you always buy like a coffee on the way to work at your local market, because you live close to one so we can like is first we get a number we like, **ugh** unreasonable. And then you can quite easily actually come up with a scenario where other data actually is reasonable. So setting sanity checks like that. Also very difficult, but important, **smiles**
19.	speaker 1	and that's the data scientist to do to do those checks?.
20.	Participant 2	Sometimes, I mean, you cannot ignore it, because then you will have to live with the consequences. And oftentimes you feel very powerless as a data scientist. Like I cannot decide how The store should set up the signs or instructions for how to put apples or you know what, but that's different than the sanity checks, I can probably influence. But then the problem is that as a data scientist, I'm good at knowing the data, but I'm not so good at knowing the domain or the field. So your key quality checking is actually that you have to have people that has the domain knowledge working together with the data people. Because it will be people who know retail who know grocery who know the stores, those people will be good at saying, is this a reasonable number or not? But they might not be so used to thinking about it in terms of numbers, they think about it more in terms of behavior. So you have to cooperate. They're like, what would do this type of behavior? mean, in the data? Yeah, and then discuss together?
21.	speaker 1	Yeah, so you would say that having an understanding of the entire process would be more helpful when doing the quality checks?

22.	Partici-pant 2	Yeah, yeah. Yeah, definitely. Yeah.
23.	speaker 1	And so our next question is about the data process. Can you describe it like from when you get access to the data that you're working with, until you're like, the data is ready to be used in the systems?
24.	Partici-pant 2	And by that, when you mean ready to be used in the system, like ready to be used in the AI model, orthe model or algorithm
25.	speaker 1	or Yeah, exactly.
26.	Partici-pant 2	mmm Yeah, I mean, often, if your database, you actually have that, ready in front of you, maybe there's like, how it is done in theory and how it is done in practice *smiles*. But in theory, you kind of have your nice data set. And then you make some graphs and you look at it a little bit things. Yeah, here's an outlier, or this is this type of distribution. Okay, the outcome is, so this is probably a model architecture, that will be good. Maybe this might work as well. And then you like, yeah, you do these things where you split the data, you use some for training, some for testing. There's a lot of like, maybe technical statistical aspects of this, but essentially, you just explore it. But maybe in practice, you will, that will be more attractive, like, it's not really a point where you have your data already in the system, like, maybe you're waiting a really long time for the data to even arrive, maybe you can even shape what the data should look like. or influence that in some way. Maybe you cannot, because it was really old, but you need some complementary data, and that you need to collect some other way. And then you need to co-operate with other departments. And, and that can be a bit tricky. And maybe you, especially when you start inspecting data. Maybe there are two some different approaches here. You could also say in AI. So in more traditional statistical modeling, you might really inspect all the data. **thinks for a second** Actually, maybe I should put it like this. In very traditional statistical modeling, which is very close to modeling, like, AI modeling, you, you want to start in theory, and you want to build an hypothesis of what your process is looking like, what your model will look like, what relationships are there, what are good predictors of someone's salary, for example, like a state, then you will sit and think and read the literature. And that's where you start in the theoretical end, like, it's probably how old you are, what you have study, how many years you have worked, etc. But then make lots of machine learning and AI, that's more the approach of, I take all the data I have, and I just throw it at the model * smiles*. And then the model will tell me what's important or not. And that can be a very irresponsible approach. Because you might easily throw in like, that's a typical occurrence of bias, right? Because you might accidentally throw in a variable, you might have hundreds of variables, you could easily have that. And you don't actually know really what is correlated or not like overnight, it will not be that one variable is just ethnicity, though, but maybe it's something that can actually give a hint about ethnicity, or gender or age or something like that. But you don't think about that. For example, if you think about buying groceries, just buying menstrual pads, that's a that's a super good indication of gender, right? If we're single households, so you will have that information capture and other data and if you just throw everything at the model, you can be

		very much not aware of that. But at the same time thinking to carefully about data and being really careful and really starting a theory and only testing this or that, that's not really leveraging the possibilities of this technology. Because the possibilities is just that you can take a lot of data and you can get something out and you can get good prediction. You don't have to do all that theoretical work. I mean, previously, of probably only research people can do this, but now it's actually someone like me, it was like a social scientist to study and also some, you know, I think I got the better derailed here. Extremely important in different ways.
27.	speaker 1	Yeah. And I think you mentioned something about influencing data. Do you? Would you say that that's a possible risk of like the data scientist transferring their bias into the data during that process?
28.	Participant 2	Yeah, probably, because in some cases, this. The this,**thinks** these biased varibles, so to say Are, are also like the best variables that you really want to use? Like, if you want to guess whether make an estimation whether someone has a kid or not, for example, then age is the best predictor for that. There is no, like doubt about that. Like if you as a human, were just to guess, if I had a child or not, for example, then maybe for me, who is actually like 30, maybe it's a little bit hard to know, could be a little bit a either or, but if it's someone like that is 40, I will guess that they have a child. And if it's someone that is 20, I will guess that they don't have a child. So yeah, most probably,
29.	speaker 1	yeah. So our next question is, can you ensure that there is no risk for bias and the data or algorithm algorithms or the systems that you use? Or would you say that there is no way of like, ensuring that
30.	Participant 2	there's definitely no way to ensuring it to 100%. But there are, that doesn't mean that there aren't things you can do to at least make it better. Like, of course, there are checks and so that you can put in place. And I think that actually something that is very key here is that it's much easier to check for biases, if you have any data on like the what I would call the sensitive attribute. So if if we say that the sensitive attribute that is stuff that is predicted by discrimination law, so maybe gender, ethnicity, age, all that kind of stuff. If you have data on that, then it's much easier to check both the system output and the other system input. Like, because essentially want to check if there is correlation to this variable. Well, first of all, maybe you want to say that we're not going to use this variable. But it's not clear cut that using it will be like, using it or not using it, it's not always the nice thing, like we can think that using it. **thinks** This depends how the data look essentially, like it doesn't feel so nice to use the under as an input, for example. But I made an example once about like predicting the the **eeh** gym like grade you get in school for the physical activity sessions. And if you base that if you try to like build a model using maybe the strength of the person, so how much they can do in like, some kind of workout like **raise hands movement**, I don't know, I know what the movement is. But then probably the strength overall would be a pretty good indication of what grade you have, like probably people have higher grades generally also are stronger, and you will find a positive correlation there. And you can kind of build a model with it. But if you were not to include gender as a variable, then all girls would get lower

		grade, and the guys would get higher grades. Because they're generally stronger. And if it's a positive model is just supposed to the relationship between them. So then you actually want to include gender to ensure the fairness. is rare you wanting to like take it out of the model. And then the second thing is that when you don't, often times in reality, you don't even have the data that you would like to check for such bias. So I have had GDPR people come and ask me like, do you use ethnicity in your model? I'm like, no, because we don't we don't have We don't have that information about customers **laughts**. But I can also not ensure that my other variables don't correlate to ethnicity because I cannot check because I don't have ethnicity data. And that's good. Like, we should not have ethnicity data we don't want to some we think is fine. Gender, we think normal is fine. And age, we think normally, it's fine to store data, but it ethnicity, religion, sexual orientation, all these things. We don't want it to be data lying around all that. But that also makes it so much harder to check up. There is no bias on that. Because I cannot check, I cannot check Is there a correlation between sexual orientation and ice cream preference? I cannot check. I will not know if the ice cream like preference actually represents sexual orientation information, the way I can with sanitary pads and like menstrual pads and gender that I can check.
31.	speaker 1	yep. So it's really interesting. I like it. It opened my eyes to a lot of new things. And I think it's really interesting. But so you would say like, it's not always the data that is the issue, but like how you use it?
32.	Participant 2	Yep And it's not like remove all of this sensitive information, and then you'll be fine. Absolutely not.
33.	speaker 1	No, like the gym example you gave That's actually like we needed to make sure it's not biased. **participant agrees** That's interesting. And then I have one last question related to data. So when you work with demographic information, is there a different routine? For how to work with that? Or do you work in a more precautious? way when you're dealing with demographic information?
34.	Participant 2	Yeah, like, then maybe have to motivate it much stronger, and really document like, why do I need this. So if you're using age, then for example, to estimate when someone has a child, you can build one model with age and one without, and you'll see that there's an accuracy difference. And then you'll use that accuracy in different brands to motivate that we actually do need age here. And so I think, in general, like documentation, and have been more thinking through like, it's good or bad to include this or not, like, it does matter quite a lot. And, like here, I don't work with a lot of in my new job, I don't work with a lot of personal data or data that is tied to individuals. So we don't have so much concerns about it, but it can there were lots of concerns around that. A lot because of GDPR. So that has that law has helped in like making companies more aware that they need to have different procedures for for that. But I think what is still lost is this connection that other variables could encode the same information. without you knowing it.
35.	speaker 1	Yeah. And that's the tough part. Because if you don't know, you don't really know how to fix it. And even if you did know, some of the things are so complex that you really don't know how to deal with it **participant

		agrees** So that were all the questions I had for the data section. X you want to continue?
36.	Speaker 2	So now we start with the bias section. We had three questions in here. And we were wondering, like, how will you work in order to that bias? Like do you have any protocols to detect bias?
37.	Partici-pant 2	If I have protocol, no, sorry, what was your question?
38.	Speaker 2	How do you work in order to detect bias? Like in the data or in the system?
39.	Partici-pant 2	so one thing is checking beforehand, before you feed any of the data that's quite related to the stuff we just talked about, look for correlations and such. But then you can also check like, the output of my model, how is it perform-ing on different subgroups? So you could see, for example, that if you have men and women, you can check, you can actually actively split up your model on those groups. And you could say that the model is performing like this or like, like this is the accuracy for men, and this is actual accuracy for women. And then sometimes you want sometimes your goal is to get equal accuracy, and you're fine with that. And sometimes your goal is not to get equal accuracy but equal outcomes. So those one case where you want equal accuracy. That's medical cases very much. Like, I don't care if men are more likely to, or women are more likely to be diagnosed with breast cancer, like that's not an issue, we just want it to be correct. And obviously, then there will be more women with breast like, and if there is a gender difference in a disease, you don't care, you just want it to be accurate. But if it's like an al-gorithm predicting someone's salary, then you don't want it to be correct. Then all of us would earn much less, yeah, not equal. So you can have dif-ferent goals there. But all of these things you can actually check afterwards and compare them.
40.	Speaker 2	Okay, so checking comparing correlation and accuracy **participant agrees**. Perfect. And so that brings us to the second question will be like, so in your organization? Are you able to identify bias? And if so, what type of bias Have you identify?
41.	Partici-pant 2	Yeah, we could see, for example, in my old job with this predicting children, which is that the model was performing quite poorly on old people, like and there because you said earlier on that you will not buy diaper unless you have a child. Well, if you have a grandchild, you will sometimes. And so. So then we could see because we did such splits like that. Okay. But I don't know if that's, in one sense, I don't think conceptually about that as bias. But it is in a sense, but but we could just see at least at the model was not like equally good at these different people. I can see in the data I work with now that I mean, now it's like a lot of data on traffic situations, and so on. And I think that that's very much these companies that are working with this day, they are collecting data, mostly from Europe, and the US, like Tesla, and all of those are doing that, too. So you can think that you're the first thing you do is not that you go to like rural Africa and try to collect information on that road. But on that, at some point, you're gonna have to, because those systems are not going to work in rural Africa unless you collect data there as well. So yeah, you can see bias in all companies data to some extent.

42.	Speaker 2	And I think now where can they look in more of a technical question in answer in the way that we have identified there is like historical data bias, or representative bias, or like measuring bias? So we are wondering, like, what do you think is like the most common, especially in the data? If it's either one of these ones or any other?
43.	Participant 2	**thinks** I think that representation is probably and they are very hard to separate. I think all our common on expense than that's the one entry really depends on the use cases where like, what problem are you solving? What data are you collecting? That in many situation I think it's representation and that can mean very different things that can be what I said now about traffic situations in different parts of the world, but we don't have them represented. It can be for for maybe a grocery company, it's more like we do we have the list data on the customers who shopped the list. So customers who shop very little they are not as good represented as and so I think always you have some kind of problem with your with really having representative data. And, and that's the same for research researcher also do not often have representative samples, and they do all sorts of things. They, you know, they stratify and they really try to deal with it in good ways. But in general, it's very, that's very hard to get, like a full, fully covered, and what would that mean? I mean, in the end, then you would just have to have data on exactly everything in the world.
44.	Speaker 2	Yeah, so that's Can you repeat the last part please?
45.	Participant 2	Know, just that maybe in one way you could maybe, maybe, maybe the representative data that is the most common one of them. If i am forced to choose one of them
46.	Speaker 2	Of course, and So now we are done with that part. And we have one question, team diversity. So we recognize that diversity plays a central role in bias attention **participant agrees**, and as a perceiver working individually. So we were wondering, that how diverse is your team? And do you think like having diverse teams can have a positive effect of on the testing bias?
47.	Participant 2	Yeah, definitely. Like I think that's that's why there were so many issues initially because there were only this young male engineers in us kind of who developed it things or a lot at least. And and you're especially when you don't have data on it, you'll be better at detecting what what is a risk for, like bias Hinden in other types of data? Like I can probably be quite good that being like, yeah, it should be the menstrual pads or you know, what, what does women buy more in general, but I would be very bad at like, it needs to do ethinicity things for example. So I think that absolutely, but But absolutely, that diversity plays a big big role here. And that's why we need to have a much more diverse feel sort of I could feel a bit lonely at at my previous job in kind of thinking about these things or caring about these things that maybe others did not care so much about if the model was better for old people or single people or so but that was what I thought was the most interesting for you go ahead and build the model I will test if it is biased
48.	Speaker 2	We need people like that, I was I was gonna say we need you in the workplace

49.	Partici-pant 2	well, we need you too, I think it's great that you are exploring this topic, be-cause then probably you will be able to also be part of that change. Yeah,
50.	speaker 1	hopefully.
51.	Speaker 2	Yes, exactly and now we get to the fun part. And so we have three cases to show you and then in the cases you will be able like I know we there is a lot of factors that to consider in order to say okay, it was algorithmic bias or it was that data bias we want to listen to your opinion of what you think so we're gonna I'm just gonna share my screen here. Let me just Can you see my screen? **participant agrees** . Perfect. So, we we deleted we had de-velop this model where we say like okay, we divided by what will happen so we put it like that data bias can happen in either the data collection or the data preparation. And algorithmic bias can happen in the development the evaluation or the post processing or deployment of the systems so if you need to like it's very straightforward but if you need to understand which one of them means what then we had like these little texts in here, otherwise we can proceed **participants agrees**
52.	speaker 1	So this is the first case I don't know if you prefer to read it to yourself or if you want me to read it out loud
53.	Partici-pant 2	I can read it and see if I have a questions
54.	Partici-pant 2	Yeah, so maybe it's actually hard to distinguish the different biases here. But I think maybe like an origin away.
55.	Speaker 2	I cannot hear you. And sorry, I hear you very, very low.
56.	speaker 1	This is strange. Can you hear me my I can hear you. I can hear both of you.
57.	Speaker 2	Give me me good. Maybe it's my computer. So give me just a second to put my headphones. Can you talk now? Okay, better? Just now. It's perfect. Thank you. Okay.
58.	Partici-pant 2	And now we're just saying that they are a bit hard to distinguish these two, I think. But I think yeah, here's a lot of data collection problem probably, like, I would just guess that they have not trained this image recognition system to on people of Asian descent. And that's maybe the mean, root cause of this. But I mean, I'm trying to think if there's other things, I mean, in the one sense, if they have not evaluated that with regards to subgroups, then it's a little bit model evaluation to that maybe in data collection, model evalua-tion, and probably quite tied together in that sense. But, yeah, probably mostly data bias.
59.	speaker 1	Yeah. So if we go to the next slide, there is one additional fact. X, could you please change this? Yeah. So here is an additional fact. So would you change your answer? Or you would keep your answer?
60.	Partici-pant 2	Not I can keep it. I mean, I think that's what I said that since Yeah.
61.	speaker 1	Yeah, exactly. So if we proceed to the next case.

62.	Participant 2	And yeah, this one is quite interesting. I know that case quite well. So here, even if you hear it's actually, in the data bias, there is one lacking here. And that I would call data generation. So the process whereby the data is generated originally, here is like where the bias steps from, because in any type of criminal records, you will kind of have an over representation of or, for example, in the US, you will have more black people in that data because the police is biased, and the system is biased and the history of the country. And that, so you have probably collected enough data, like your maybe collected all of the criminals, maybe you have perfect data, but there's still bias because there's bias in the world. So that's, to some extent, this becomes a philosophical discussion about to what extent data is a good representation of the world. So either you could insert one here, maybe this is not the question from it, but but it's not the collection process, that it's a problem. It's not the preparation process. It's like the generation of this data in the first time that they start.
63.	speaker 1	And the societal issues may be that we have been struggling with
64.	Participant 2	it maybe to some extent, you could blame it instead then and say that okay, it's then it's not data collection is not data preparation. So then it must be some algorithmic bias instead, you maybe you could argue like that instead then that the model development has been done with the wrong variables,
65.	speaker 1	yeah, yeah. If we proceed to the next slide, there are some additional facts, especially the last one, I think, is what you brought up as well. With higher police presence and such neighborhoods, which will lead to more arrests. So as you said, the data is correct, but it's not like used in a good way. Or representative. We can continue to do last one, maybe.
66.	Speaker 2	because my headphones. I didn't hear the last part. Should I change now?
67.	speaker 1	Yeah, sorry. No, it's so this is our last case. Yeah.
68.	Participant 2	Yeah, this is probably similar actually to case number two. Like, because of like, historic bias in who can be interested in technical things and Who can study this? This or that much math or whatever, there is probably historical, there is probably a data that women are less likely to open some of these ads. But that does not mean that we want them to be shown less. So it's again, it's kind of like data data generation, you could almost say. But but you could probably have the same thought I had there about maybe that's what we should call algorithm bias and that you have not kind of taken that into account when you try to develop your model.
69.	speaker 1	Yeah. If we proceed to the next slide, there are some additional facts here as well. And this was actually, they couldn't really determine what the reasons were. But here are three possible reasons. Well, which one would you choose if you had to choose out of these three?
70.	Participant 2	I think that the the two first are actually the same. Yeah,
71.	speaker 1	yeah. Yeah.
72.	Participant 2	But could probably be the third as well, I've not really thought about that, that maybe there's higher competition for women's attention online. So to

		say that women are more so subject to advertisement in general, because we are subject to so much like clothes and makeup. But I've not really thought about that before. So that's kind of interesting. But probably the first student like, I'm guessing that women were still less likely to click on that, like, don't they? There's a point in, ignoring that. But then that has reasons. And just because it is like that we shouldn't just like comply with that and leave it like that. Because that's, that's a problem, which is
73.	speaker 1	that's an important thing to remember that just because it is like that doesn't mean it doesn't have to change. It needs to be changed. Yeah. Yeah. That was all of our questions. And do you want to add something else? Before we wrap up?, do you have anything else you want to add or ask?
74.	Speaker 2	Not so far, I'm okay. I was just like writing just in case. The last part of the recording is not that good because of the headphones.
75.	speaker 1	Yeah. Yeah. Thank you for taking the time to do this. We learned a lot. I know I learned a lot. All right.
76.	Partici-pant 2	Just don't forget that it would be great if I saw that they were anonymized with regards to me. But it's also good if you can anonymize and not write X for example, big Swedish company or something like that?
77.	speaker 1	Absolutely. Absolutely. We will do that. And we will send it out to you so you can read it. Yes. Yes. Thing
78.	Partici-pant 2	Good luck Reach out to me if you if you have a question about something or you know, if if you would like to talk about it more, I'm very, like passionate about this. Yeah.
79.	speaker 1	Very exciting stuff you're working with.
80.	Speaker 2	All right. Yeah. Thank you so much. Thank you. Have a nice afternoon. And enjoy the nice weather

Appendix 10: Transcription - Participant 3

1.	Speaker 2	Perfect. So I can give you like, a little presentation of what our research is about. So the title of our thesis is bias in the context of artificial intelligence systems. And so the focus is to tackle these from a data point of view. So we're interested in finding out how the data affects the outcomes of the systems and identify where in the process the bias, of course, so we're thinking like, it could be in the data part, or in the algorithmic part. But of course, we understand it's a little bit more complicated that. Yeah.
2.	Speaker 1	So yeah, before we continue with this, I want to go through the ethical things, you have the right to cancel the interview whenever you want. Or if there are any questions you would like to skip, that's okay to just say it, and we will skip it. And then, with the help of the recordings, we will be transcribing. And after that, we'll send it out to you. So you can read through it. And there you have the chance to change or add or delete anything you'd like. Okay, so now we can start with the questions. We have to warm up questions just to just to get to know each other better. So the first one is, could you give us a quick introduction about yourself?
3.	Participant 3	Okay. My name is X and surname is X. I'm from Turkey, and living in Istanbul. I'm X years old. I'm working as a data science lead right now. my bachelor's and master's degree are on statistics. And also I'm studying on PhD on statistics, then I'm focused on AI and machine learning algorithms. Mostly.
4.	Speaker 1	Yeah. Could you tell us about your experience working with data and artificial intelligence?
5.	Participant 3	Okay, actually, since my university, I've had some projects about artificial intelligence and machine learning algorithms. Most of them are about the direct discovery gene, the human gene, researches mostly on about data driven or machine learning perspective. And also, after the universities, I started to work as a market research company. In basically we collect the data, like, you know, the marketing service surveys, like the products, it can be about the services, customer satisfaction, something like that. And we analyze the data in according to our, the service responses, and we just tried to understand the customers the behaviors, extract some insights from products or customer satisfaction. After that, the right now I'm working as a data scientist, I'm most mostly try to understand the customer's behavior as well. I mean, you know, we need to find some patterns, we need to predict customers next steps, it can be about upsell, cross sell. And also credit limits, we just are finding out the patterns of customers, they will likely to do that or not.
6.	Speaker 1	That's interesting. Thank you.
7.	Speaker 2	I can see that you had like, very good experience, like, I can academically and then as well in practice. So this is great to get to understand more about the whole process not only for from the theory, but only as well from what you have worked with. And so we'll start with the

		questions on the artificial intelligence view. We only had two questions. And the first one is the way that we define artificial intelligence And we define it in more of a social way. So the definition is AI is the ability of the system to identify, interpreted, make inference and learn from that to achieve determinate organizational and societal goals. Do you agree with our definitions? Or will you change or add something to it?
8.	Participant 3	In general, it this definition, I think it's okay. But sometimes it can be changed depends on the situation, especially interpret and making inferences, you know that some cases we cannot interpret about the problem, we just need to find patterns and follow the rules and predict the problem solution. So not always interpret and make inferences, just I want to say about it.
9.	Speaker 1	Thank you. That's a good insight.
10.	Speaker 2	Yes, we are gonna use that. Thank you. And, and the second question is, the literature that we have read points out that there is a need for addressing the dark side of artificial intelligence, and such as the race for the system, acting bias. So is this something that is talked about in your workplace or among your colleagues?
11.	Participant 3	Yes, we talk about the dark side of AI, with my colleagues, but it's mostly about the, you know, psychological perspective. Because I believe that the people are kind of broken. I mean, you know, just AI, learn from data, or find out the patterns from the humans behavior. So, sometimes, it can cause some dark side effects. Using AI. Yes, yeah. Yeah,we talk about it.
12.	Speaker 2	Okay. And do you also, at work? Do you do something about it, or its use among colleagues?
13.	Participant 3	I don't work about the dark side of AI, just you know, is conversational things.
14.	Speaker 1	Okay, so if we continue to the data. Our first question is, when collecting data, do you have specific requirements that you follow or other other things you try to consider when you collect your data?
15.	Participant 3	Actually, we don't have any specific requirements for collecting data. Because the data just come from or the system, you know, that just like I mentioned, I work at X, and data just are produced by the customers. So we don't have any requirements for collect data is it's automatically produced from the humans to our customer, something like that.
16.	Speaker 1	And you use every data point, there are not certain things that you decide to not use.
17.	Participant 3	No, we use all of them.
18.	Speaker 1	Okay. And then that brings me to my second question, how do you make sure that the data you use is of high quality
19.	Participant 3	it's actually we have a big teams on the company some of them come from engineering parts data engineering parts, actually, they ensure the data have high quality or not using some tools such as you know, you can analyze the data is for example, just I want to give an

		example about it. We think about, just think about it age, the age can be between in some continuous numbers you cannot see the customer age is 200, it is not impossible. So we need to check the some kind of rules. Like, you know, the minimum value maximum value, the average, something like that we just use this kind of metrics to see if data is true or not.
20.	Speaker 1	Yeah, that's interesting. If you have any other examples, feel free to add them as well. These are exciting stuff, and I think we can use them
21.	Participant 3	I don't know right now **smiles**
22.	Speaker 1	No, it's okay. So could you describe the process from when you get access to the data that you're working with until it's used as input for your systems?
23.	Participant 3	Okay, just I mentioned, we have a data engineering team, they collect the data, too, to our big data clusters. So I can use the data from from that the clusters. But all of them are raw data, I need to aggregate this data like in the based on subscribers, it can be about the specific time periods. For example, last one day, last one week, last one month, something like that, I need to aggregate it. So just we need to, we need to *mmm* determine a problem. After that, we use some correlated data about what I try to solve about my problem. Just it's all about, you know, the user guided decision. To check if this problem can be correlated with data or not. But we use all of data to find some patterns. But we don't have any, just specific requirements, should I use this kind of data information? Or not? Just I ensure the data is correct or not. That's all.
24.	Speaker 1	Thank you. Um, then we have another question. Can you ensure that there is no risk for bias in the data and algorithms and the systems that you use and work with?
25.	Participant 3	To be honest? No. Because, as you know, a company, they just want to make profits and revenue. If my data provides, provide us with more accuracy, we don't need to **thinks**. I mean, we don't have any requirements about the bias. Just we use the data. Does it give me more accuracy? Or not? Does it give me more profits or not? And we use them?
26.	Speaker 1	Yeah. A follow up question. Do you think that anyone could ever ensure this? Or if anyone makes a promise like this to ensure that there's no bias as a system? Would they be wrong to do that or not? It just seemed like the company context. Like overall, do you believe that? We as humans can make that statement to ensure that there's no bias.
27.	Participant 3	Honestly, we need to ensure there is no bias about our data. We know that data just mentioned people are broken. So we need to ensure all of them are is accurate or not? . Just not from data, and also from the algorithm perspective. ,
28.	Speaker 1	Thank you. And this is our last question and I'm before this section, do you work in a more precautious? way, when you're working with data that contains demographic information, personal information?
29.	Participant 3	I believe that we don't have some more privacy of our, the demographic data. I mean, we don't use the data that come from the

		religious particular idea or ethnicity, we don't use them, because and also, we don't collect them as well, actually. But we use some data. With demographic information, like the gender, age, education, maybe it can be about the location. We use all of them.
30.	Speaker 2	Do you treat this differently based? Because of them being demographical? Or?
31.	Participant 3	Yes, because algorithm learn from data? So It can, make prediction From, you know, this kind of data at age gender?
32.	Speaker 2	Yeah, so I can continue with the bias part. And so we were wondering, how do you work in order to detect bias? We know, you might not do that. But maybe you have an insight, if in your company people work to the debt buyers or not, as you say,
33.	Participant 3	we don't detect actual bias. But of course, there can be some protocols to follow them. For example, you know, we don't use some data, more private, like the religion at the city. It can be about political views. This kind of data is really dangerous when you use as giving input to your algorithm. But actually, we don't have some protocols about this. Detecting bias
34.	Speaker 2	And so, based on your experience, . So what is the most common type of bias? So we want to have more of a technical approaching here. So rather than say gender bias, we would like you to maybe think about historical bias representation, bias measurement, sampling, or bias on the algorithm itself.
35.	Participant 3	Actually all of them are very important. Detecting this type of bias. Because you know, the data come from the historics. And, and also, sampling is very important. If you have a very small sample that is present for a big population. It doesn't work, actually, it does represent your population. And also something is very important as well. Algorithm. Yeah, algorithm, sites, algorithm perspective, when I think about it, it's also important, but it's small effects. I believe that is it's the most bias types come from data.
36.	Speaker 2	Thank you.
37.	Speaker 1	Happy to hear since that's the perspective we're working with. **participant smiles**
38.	Speaker 2	Yes. And so one last question. And we recognize that team diversity plays a central role in finding bias, and opposite of working individually. And so we want to proceed with the following question, which is like, how diverse are your teams? And do you think having diversity in your teams have a positive effect on the team bias?
39.	Participant 3	Actually, diversity is very important in our teams. You know? If you mean about religious ethnicity. Because if we don't have any experience, because of our ethnicity, or religion, we cannot understand or we cannot predict how it cause kind of problem. Because we don't have any experience, if you have more diversity in your team, they can understand it, they can share prediction about collected data. And using some problem type of algorithm. It's more, it's really important. It's really important diversity. ,

40.	Speaker 1	thank you. And now we have three real life cases that we would like to present to you, and then we have a question for each of them. So if you could share your screen, that would be great.
41.	Speaker 2	Perfect. So we have identified that bias cool happen in two different ways. It could be the data bias, that it that we decide that it might be the data collection, or the data preparation. And then the algorithmic bias could be different ways. The model development and model evaluation, post processing and deployment, this is where we have identified but if you have any insight to this, you're welcome to add. And then of course, we understand like, maybe algorithmic bias can be part of that bias at the same time, because they might go together. And so with this, say, we want to present you the cases. And then in the case is you can say where the bias will happen. And you don't have to go that deep into picking each one of them. You could pick either that our algorithmic, and then you need to understand what they are. And then we had like, here the information. So once you're ready, we can go to the next case. First case,
42.	Participant 3	just give me one minute. Okay, let's continue. Case one.
43.	Speaker 1	This is our first case. I don't know if you prefer to read it to yourself, or if you want us to read it for you.
44.	Participant 3	Please, can you read the case, please? Thank you.
45.	Speaker 1	Yes. So this is the first one that's about a man of Asian descent that wanted to apply for New Zealand passport online. And the photo he uploaded was rejected. And that was presented with an error message that said his eyes were closed and that the photo did not fulfill the requirements of the system. His eyes were open. And after three attempts with three different photos, he had to reach out to the passport office, and the office blamed it on bad lightning and shadows in the eyes. So this is what happened. And then our question is starting from the data collection, to them model deployment of the system, where would you say that the highest risk of bias occurring would be for this case? Where do you think that the bias happened? Ah,
46.	Participant 3	actually just start the first part of data bias. Because just I assume that in New Zealand passport office, they don't have an historical data come from, you know, Asian descent. If they don't have any about sampling of this image recognition data. The algorithm cannot detect properly actually, because we because, you know, Asian descent people have, the more closed eyes, if the data don't have this kind of represent of Asian descent data, algorithms cannot find actually the property, data collection is the very important in this kind of cases when I consider about data preparation is also important. But the analysts they don't know about the this kind of data includes in their population or not actually. So, you know, there are many possibilities about so they cannot check all of them.
47.	Speaker 1	Great. So if we go to the next slide, we have one additional information section that says that there was an under representation of people of Asian descent and the data set used to train the model. So, I guess that's what you were saying as well. **participant agreess **

		So, you don't want to change your answer, because this gives you an opportunity to change your answer based on this additional fact. But I guess you're going with the same answer.
48.	Participant 3	**participant agreess** Yes is the same
49.	Speaker 1	thank you perfect. So this is our second case. And it's about an algorithm called compass that was used in the United States to determine whether or not a defendant that is awaiting trial will reoffend. And the decision is based on more than 100 different factors, such as age and criminal history. And when analyzing the results, it showed that dark skinned people were more often classified as high risk among and the defendants that did not commit new crimes. Yeah, so once again, this is the case and we're wondering, Where do you think about your secured?
50.	Participant 3	the collect some data come from? I mean, some privacy data, like dark skin, as well, in that 100 different variables like hate and criminal history, but also, I think that the feature, one of the feature can be about dark skinned about the skin color. So, data collection, okay. And also have some problem, but mostly about the preparation is the most critical effects. Because in historical being of that, the government or people, I don't know, how can I describe it, they treat in a Bad way dark skinned people, they just put them jail. Maybe it can, without them being criminals Maybe we don't know because it's very bias. In historical perspective, I know that. So data preparation, you need to eliminate this kind of inputs, like not skin color. Maybe it can be about the gender or something like that. So that the probation is that I mean, data bias have the most effects about finding high risk classifieds for dark skinned people
51.	Speaker 1	Thank you, then we also have some additional facts. So the first one is that the granell granularity of the data was different. And the second one says that it use the prior error status of the defendants families and friends to determine their likelihood of them very different reoffending. And then the last point is that not it is not unusual for minority communities to have a high police presence leading to a higher number of arrests, drawing the conclusion that the residents of minority communities have a higher number of dangerous people due to having a higher number of arrests is misleading due to the factors such as the police presence. So knowing these facts, do you want to change your answer? Or would you still go with the same answer?
52.	Participant 3	I still might as well ask for it is well, it's in this information data. Yeah.
53.	Speaker 1	Okay, then let's continue to the last case. In 2018, it became known that the ads related to steam carriers were less likely to appear for women than men. The ads were run on many big different social media platforms such as Facebook and Instagram. However, women were less exposed to the ads, and they were appearing more frequently to men. Why do you think that is? Where did it by a secure?

54.	Participant 3	We know that in STEM careers that we consider in the historical perspective. Again, the woman was not allow on science. So the historical data were are very bias when we consider about the man or woman careers in STEM. So if the algorithm learn from this kind of historical data, because it's also unbalanced data. *eeehmm* In the STEM careers, I don't know, actually, the share of the balance is obvious that there are unbalance when we consider the woman and man In this distribution of in STEM careers. The mostly the problem about data, not algorithm.
55.	Speaker 1	Yeah, yes. Okay. That brings us to the additional facts. So the people who research about this had two possible reasons for this. So we just want to get your opinion on what you think. Because they didn't they don't really know the answer, either. They're just elaborating on possible reasons. So the first one is that the algorithm learned the pattern from the consumers and women were less likely to click on the ad, therefore, it's trained itself that showing it to men would be of higher value. And then the second one is that presenting ads to women costs more and advertisers must pay more for exposure to women. And the algorithm is programmed to make the most cost efficient choices and is therefore leaving out women. What do you think about these two?
56.	Participant 3	Okay. In this problem, they also manipulates the algorithms. So **thinks** I think in according this additional information, algorithms have more effects in this kind of problem. I think so not on the data. Of course, data is important. Maybe I can say equal effects
57.	Speaker 1	You would say that these two could be possible reasons as well.
58.	Participant 3	Yeah. Yes.
59.	Speaker 1	Okay, thank you. That was are all for us. We were wondering, do you have anything you want to add? Or Speaker 2? Is there something you want to add?
60.	Speaker 2	Um, no, I okay. But if then is do you have any maybe advice or anything else you would like to say about the research or the whole conversation?
61.	Participant 3	Actually, I really enjoyed about this research. Your hypothesis and your research are very valuable. And it's worth doing that. I hope your search results can happen what you expect. **smiles**
62.	Speaker 2	Thank you.
63.	speaker 1	Thank you. Yeah. Taking the time to do this with us. It was really full of insights for us. I'm so thank

Appendix 11: Transcription - Participant 4

1.	speaker 1	When the transcription is done, we will send it out to you. So you can go through it and delete, remove, add whatever you want. So our first question is a kind of a start a warm up question that would is, could you provide us with a quick introduction of yourself?
2.	Participant 4	Yeah, sure. So, I, my name is X, I am a machine learning engineer at company X I hold a PhD degree in computer science, also have a master degree and the Bachelor of computer science. So I work in before in telecomunication projects related to machine learning and sometimes not related to machine learning. And my my main areas of research and development were time series analyzers, fault management, software defined networks, NLP. And yeah. Basically, that's the quick overview, I guess.
3.	speaker 1	Yes. Thank you. Could you tell us about your experience with data and artificial intelligence?
4.	Participant 4	Yeah, so I work with the with data, pretty much, for all of the Masters and PhD and also in at my workplace, in the different companies. So I started working with data in machine learning around 2010 - 2011. When I developed it, some it was before TensorFlow and before all of these libraries, like implemented the self organizing map, it's also called kohonen map. So the tree knowledge to understand and analyze the behaviors, it was a DDOS detection method. So basically, the data was networked traffic volume, some data, some features that we extracted from this input data. And then we use it Kohonen map to aggregate what was regular legitimate traffic, and what it was DDoS attacks, so like to, to understand these different patterns. And since then, I worked with, like, I haven't thought a lot about it like to have these big categories. But basically, it was network data. Most of it was times temporal data. So it was volumetric and different database, but these were described timestamps. So most of what I did was related to that. And I also had the, like, I worked in with data, half a it was non human generated. So it was like, basically, data from machines and from industry and from like, different phases of the production. And then another part was generated by humans. mainly on the NLP project. And we also worked on a bug report, text classification. So basically, it's given a bug, what is the development chain that should handle this bug? And in other other, like other projects, as well, like, not say the details of them right now? We can discuss more later. But yeah, basically,the human generated data was more about it's more linked to the NLP models that are using and yeah, right now as well. It's a lot of human generated data.
5.	speaker 1	Thank you. Great. You want to continue speaker 2?
6.	Speaker 2	It's really interesting, because yeah, you do have like, a great background on academically as well as experience. So yeah, computer science, but then you also had like hands on experience in the industry. And that's something that we really enjoy, like to actually get to know your academic experience and as well Your experience at work. So I'm going to start with the AI section. So we had two questions. So the first one is the way that we define artificial intelligence, our way, of course, artificial intelligence really hard to define and our ways a little bit more social. So we would like to give you the definition. And then if you have anything to add or change, then you can do it. So the way we define it as AI is the ability of a system to identify, interpret, make inference, I learn from that to achieve or terminate

		organizational and societal goals. So do you agree with this definition? Or would you add or change something? 7 Let me copy in the chat.
7.	speaker 1	Yeah, we can send the chat maybe?
8.	Participant 4	Yeah, **participant reading** the ability to identify, interpret, making a difference and learn from data to achieve better than individual and societal goals. Yes. For me, it's, to be honest, even though I work with AI for a long time, I always it does feels like the AI definition. It's a moving target. Because we look like at some decades, decades ago, we didn't have that much of machine learning and etcetera. But it's, like, if you had a lot of ifs a lot of right If This Then That if this than that two, were basically transferring the expert knowledge into the systems. It was a call it AI, but right now, a lot of people say that's not AI, because so AI, it's it's more of a philosophical discussion, because basically, this definition and a lot of the AI definitions, they always, like they can be applied to a lot of computer science work, basically. So if you build a program that can be able to like, let's say, to calculate the body mass index, BMI, in English. So basically, if you say, your age, your height and your weight weight, then you can take a this is your, like, your healthy or not. These, I mean, it's be able to identify, interpret, and make an inference and learn from data to achieve a goal. But it's but it's not, it's not what most people would call it AI. So if you can have like, get new data processed, and then have a different result, this like for Based on this definition. Some people might argue that these AI,
9.	speaker 1	so how would you describe it if someone asked you what AI was? So how would you describe it?
10.	Participant 4	Yeah, I mean, AI is It's more like it's a moving target. So it's like this area, that it's always pushing the limits, pushing the ideas about what computers might interpret. So the, like, the current limit, the current change of knowledge about, like, right now is way beyond what we had 20 - 30 years ago. Yep. So right now AI is, oh, we should have a car that's able to drive but like, let's say 10 years ago, 20 years ago, we could say, hey, we need to have a car that is able to drive itself. But right now it's more like it should make decisions about how it should behave and how it should interpret people interpret different inputs and not kill anyone and etc. Like even inserting some ethical decisions into the car itself. So yeah, it's I get that. I mean, it's, it's always a slippery slope in when we are defining the research areas. So for me it's about that edge about that. It's it's one of those fields where it just keeps pushing its its limits and its boundaries. Over the decades.
11.	speaker 1	Yeah, definitely.
12.	Speaker 2	Yeah, that and that is completely true. Like Many years ago, we were thinking things that we're doing at the moment is AI, like automatisation, and so on. And then now because we are achieving to do these things, it's like, no, we're not calling it AI anymore. Now something else in the future might be AI. So it that's a great point that you're saying, like from the past, and right now, the definition just keep changing. And that's the thing is quite hard to define. And we have a second question. And so the literature that we have read points out there is like a need for addressing the dark side of artificial intelligence. So these darts lie is like risk, such as bias. So we were wondering, is this something that is talked about at your workplace and among your colleagues as well?
13.	Participant 4	Yes, it's either, that's why I, I, I worked with both sides, like I said, like, more human facing models, and more like machine facing, but it's always impact to humans in the end. And even for my PhD, it's always like, has

		this impact, but I want these different degrees. Yes, it is something that we talk about is this, it's something that we talked about. So I can say that there are, there are two aspects of it, about the the, let's say, the dark side, or the ethical part of it. So that's what I realized, like, that's what I that's my take on it, it's that when you are in research, as opposed to the industry, you have more freedom about that. So unless you are specifically thinking about like, the, in this research area, this is not you might be a little bit more loose in the sense of these ethical aspects. So for example, when they're calling about were talking about effective computing, which will, which will, for example, understand the impacts of social networks on another center, they use it behavior and how it might be related to the pressure and some mental disorders and how it factors etc. So you have more freedom to analyze it and to see and it is a discussion but because it's research, you might have like might losen the restriction a little bit but if it was not research which was more about production, then the stakes would be higher because it might be affecting millions of people. So etc so in research there is a difference and in the industry as well it's it's the there's this difference when you are like looking at, in like a factory data about cell phones and this stuff, and when you are looking at something that we will, like, affect users at the more immediate level. So it's, there's the I maybe because I am very aware about this. So have the for example, if you are doing developing computer vision facial recognition algorithm, that's something that I take account I take take into account about our data set we'll need to have into account the gender representation, the ethnic representation the like the of the even the age and like the any like about people with disabilities so yeah, I guess it's it's something that I take into account. So when I am closer to should these to user face your human faces, it's something that I take into account So yeah, I can say that it's not the top one priority. But it's something that we take into account.
14.	speaker 1	Great, thank you. So if we move forward to the data part, we have some questions regarding data collection and how you work with data. So the first one is, when you collect data, do you have any specific requirements? And that you follow or any aspects you try to consider? How does the data collecting work?
15.	Participant 4	Yeah, I am very. Like, I'm very annoying about this, because of my experience and the research and like, because
16.	speaker 1	that's good, we want to hear your experience.
17.	Participant 4	what I see a lot of younger developers and researchers is that they focus a lot on the models and the approach and the techniques. But when I am working on a project, and because I am also the tech lead for the team, and there wasn't the best in different teams, I always, this is the most important part. So this is the part that we need to always keep in our mind. So again, I can say that the there are some interface, some some interlayer interlayer communication that we always need to make sure that it's working. So the first thing is to understand the domain problem. So to understand what is the real problem that we are trying to fix? In the industry, it's in the end, when we are discussing about metrics and this stuff keywords, etc. What do we need, like the most important metric in the end is money. So, yeah, how we are going to translate those metrics into business value. So these days, we need to have this clear interface, this clear definition from the problem that we are trying to solve. So I can just make up some problem. So let's talk about the retailing, for example, you know, you want to make a

		computer vision or algorithm that will say that will sell that will send the targeted advertisements to different personas and different segments. So in the end, the metric that you count is how much will be converted into sales. So basically, you need to convert, you need to show aligned to match this with the data. And then we need to not look at the amount of data, not necessarily having a lot of data, but we need to have a data distribution that is that match what we will see in the real world. So for that specific use case that we have in place, so if you have, let's say, a jewelry store, it might you need to have data input that will represent what you see in real life in a jewelry store. So it's very different from a chocolate chocolate store or from a clothes store or something. So it will have these different data distributions. So one thing is that okay, what is the end goal? How what, what is the business value where it's the money that it will generate? And then how this goes back to the data and what is the data distribution that we need to map to that we need to to ensure. So basically, the trend is to tie the whole thing together, and then coming for the data distribution, make sure that it's it's well represented. And then from that we can move forward with the model and with the validation itself. So if we have, for example, an anomaly detection problem. It's it's something that I have work in the past and one problem that by definition, if you want to detect an anomaly, you are looking at unbalanced datasets. So some By having these clear definitions about, okay, we are going to have these an imbalanced data set. What is a positive example? What is a negative example like these, these definitions need to be very, very clear. One problem that I saw with younger developers that, for example, they they, in anomaly detection, they use the So, let's say for me to detect the default, the default possibility of some. So, that case, I, what I saw in the past was that they consider the because the well functioning device might be a good thing and a bad bad function divided by might be a bad thing, they consider the well functioning a positive example, and the bad functioning negative example. But this is completely off, because it will match your procedure on your recall your f1 Because what we're looking for are the fault device you need to this is the annomaly that you want to detect. So, for example, the like, then you need to look at the data that you look at. And if you're going to have these, you need to make sure that for example, the all of the even though you have this big class, well functioning device, you need to have data diversity in that class, so that to you, for example, it doesn't it doesn't make sense to have like a giant data set where I'll have a lot of samples in the function that are very similar in its features, when you might have a very different data distribution, because even these be classes, they might still have those different so it's likely to represent this in the data itself, I can not represent but at least make sure that you have the data diverse. So basically, these is the the main thing that one of the example that we need to take care and to take into consideration we are building these datasets and looking at the data. Thank you.
18.	speaker 1	Our next question is how do you ensure that data you use is of high quality? I guess you've talked some about it, but is there anything else you want to add?
19.	Participant 4	Well, we basically as I said, I the main my main concern is about the data distribution, the data representation, we need to and you can only know if the data distribution is correct if you understand about the problem and about the use case that you will need to support and the how how can you ensure this how can you ensure that you get the data distribution right? It's that you need to have these research questions for this goes for any problem basically. So the okay we want to send targeted advertisements to different personas, different groups. So you have these research questions about

		okay, how does the location of like you need to come up with these things like Okay, does the location affects what it's better to advertise? Because the gender affects what it's better to advertise the age effect and then you need to set up the data so that you can isolate those those different factors so that you can answer those questions when you run and when you look at the experiments and even in the end even though you do a lot of these analyzis all of these you need to validate with the real world data and when you're validating these please the question that I always do in the like job interview. Okay, like you I usually I walk the candidate through the through a machine learning problem modeling and then I one of my final questions is okay, you did the model, it was good 90% of procurements etc. But when you go it goes to production. What like it's the Oh, it doesn't get anything, right? What happened. So basically, this can always occur even though we take these measures. So it's about looking at the real world data distribution, see how that match with what we have. And so either it's completely off, like those, or we have some of the some use case that we talked about before it's misrepresented in the real world data. So we have a lot more of the so if it was not, let's say, a jewelry jewelry store or something like that, we the demographics of that would be very different from fast food. Restaurant, for example. So basically, it's in the end, it's always comes to validating it with the real world data. Just one last thing is about when this goes wrong, is that one thing that should look back to the data inputs and see if there is something else that you are missing. So for example, I got like, a supervisor, these master's students where they were having, like, when I started working with them, they were already working on the project for some months. And one thing that I realized is that they, they were getting very bad results. And then I even like talking to them realize that the problem is that they were not taking the, into account the temporal data. So it was anomaly detection, but they were just looking at the data points, and not seeing the correlation between different data points in time. So for example, this is the feedback loop that we we need to do to understand the whole process and then okay, looking back at the data, is there something else that we can add to it? Is there some analizes ? So? Yeah, it's, it's like, the quick and dirty feedback loop
20.	speaker 1	Thank you. Yeah, we've heard a lot about the correlations and we've learned about the importance of being aware of them and analyzing them.
21.	Speaker 2	So I have a follow up question, actually. And so I know you are in the other side, you are in the machine learning engineering part, like so I was wondering, like, do you work closely with the people that collect the data? So you both I know you're saying like, you need to understand the requirements of the what is the problem? And then you use that in order to use a specific data and but then, like, do you work closely with like data collectors data scientists? Or do you use like wait for them to give you the data and then you from there to work?
22.	Participant 4	Yeah, well not because I'm the tech lead, I always I talk to the stakeholders directly, like with the people that collect the data, etc. But another part of the team they don't talk with them that much. So I am more of an interface. But this is a even though I talked them this is a problem in not only like it's not a team specific I can say that now places that I work in before we have these problems. So we have this gap between the the business people and like the the teams, like the different teams and how they need to communicate knowledge so that we can have okay, this is a correct it's a proper data set, this represents the problem good way because the tricky thing is that

		sometimes there is nobody that has like the whole information. So even the like the and they are not talking about the company that I'm working for right now. But what you might want to have a lot of times is the business people know spots of the of the system, but they don't know about others in they have of some misconceptions, and the operations team, it's the same thing. And then you have the machine Learn team and sometimes you have never even like other teams in this picture. So I guess, the human knowledge, it's the gaps in the human knowledge, it's one of the problems that leads to misconceptions and leads to the data distributed. So it's not like, Okay, we have this perfect tech tree and specific documentation, and we can just do the, it's just a one to one mapping that we need to do between these two things. It's a lot of more gathering and asking around and investigation like just translating from one system to another.
23.	speaker 1	So, our next question is the pro what does the process look like from you getting access to the data until you use it as input for your models?
24.	Participant 4	So, phase one, the definition what is the problem what what is that we need to solve then Phase two would be to mapping the data source or like all of the possible data source that we we might have and then from that, we want to understand the possible model for visit a time series analyze model, NLP model, graph representation for a new we will use ref neural network etc. So and from also we have this show we can Okay, so, what are the data mapping that we need to do? So, usually, I had like, two different phases. So, they are after the mapping of the data source. So, one of them is having the, like the these abstractions or these data representations on the more lower level on a lower level, like it's a few attacks. So, messages or if you have you, were talking about marketing campaigns, for the advertisements, so, okay, we have a marketing campaign itself, and then you might have the individual advertisements and then have the personas, trying to identify these abstractions. And then there's another layer where we have the process the pre processing entities. So, for this revenue network, I really did representation for for this simple tabular classification model have been another presentation. So it's like having the possibility so yeah, I would say that it's clear these four steps problem definition mapping of data source basically abstractions, and then the pre processing the processing for the input for the model inputs.
25.	speaker 1	Thank you, we have two more questions and then section and the first one is would you like you as a team or you as an organization, say that you're able to ensure that there is no bias in your data or in your systems?
26.	Participant 4	Yeah, I mean, then we have to go back to the definition of bias. Like for sure that we have some bias and a lot of the companies who have bias in comparison with like, the more data on the worldwide like or some anything that's outside the company, so if as I said, like different stores, different places, will have different bias in its data. But I guess that the the important question or the like the How can I say like the the main thing is, we need to be aware of that of this bias and how it will affect so it all data will have will have bias, but we need to be aware so that we can Okay, this is how we are going to like is this something that we need to take into account or not take into account and it's the like for the both ways, because sometimes we need to have this bias because our data distribution has this bias and we need to keep these with the to the to the forward propagation of is these bias to have everything right. But sometimes you might, it might be a problem because you have these. So, for example, just imagine that you might have a surveillance system that will just track user like, people inside the store will pay you store. And like, say File trigger, in case it detects some, some suspect behavior. not sure just look at their keywords or their

		score, f1 score, etc. But also look at the ethical parts of it. So that means So, okay, if we have, like, we need to, we have some bias, but we don't really to avoid racial profiling, or gender and profiling are these are these types of, it's more about like, Okay, we know that we have bias in our data, but with this nature, it's not a matter of removing the bias, but taking the bias into consideration about whatever we need to do. Yeah.
27.	speaker 1	Great. Thank you. And then that brings us to our last question. Do you work in a more precautious? Way in relation to bias when you work with demographic data?
28.	Participant 4	Yes, yes. It's, as I said, Yeah, actually, this question is tricky, because I'm not sure if somebody will, will answer. No, I don't like about it. But it's, it's very, it depends a lot on the on the downstream tasks, like on what it needs to be achieved. So as I said, if you want so to advertise for like, older people, not specific, like, I'd say, male middle aged map, the person for example, so okay, this is a bias that we need to take into, like, we want that bias that sense. So it's not like we need to clean it. But it's, it's basically, instead of like, we, I, we might using might even liberate these demographic knowledge to to achieve our goal. It's also like, be careful about not do any, like, unethical choice
29.	speaker 1	Thank you., you want to continue?
30.	Speaker 2	Yeah. And that last point that you mentioned, is true. Like we write about it in the in the research, that is like, as long as you're not discriminating somebody. So for example, if I do in an advertisement for a job that I know, older people might not be able to do. Then of course, like there is like, especially in advertisement, like you have like a specific age that you click on. And so now we just got to the bias view and the bias start with the topic of bias. So we have like three questions here. And one more thing is like, how do you work in order to the Detect bias in your data sets? And I know AI, bias is tricky, but let's say like discrimination bias, like things that will discriminate people.
31.	Participant 4	How can we avoid that?
32.	Speaker 2	And how do you work to detect does bias?
33.	Participant 4	Oh, yeah. So it's, it's linked a lot with the problem definition, what we know about the, like, the input data, and I. So for example, sometimes it's invisible to actual to us, we can't know that. So for example, if you receive credit, I'm trying to just like to disclose any anything that I like to specific about projects I worked on. But sometimes it's very obvious, like, you make sure like, you need to set these things up in the in the input data, so you can look at the, at the just the datas distribution and the results that you had. And okay, we can see that we can show, for example, do a feature importance. And then we can detect, oh, this feature that's closely related to gender or to age is having a high impact on the final results. So this is we are having a bias towards that. So even though we have these good results, which look into these as well to make sure that it's not just a dataset bias that we have. So these are, like, it's somewhat, in some cases, very obvious. These because for example, if race is a feature that you are using, then you can just look at it directly. Sometimes it's indirect. So it might be just a computer vision, detection face recommendation, and you need to do what you need to do is analyze them. Understand, okay, but what are we getting wrong in our algorithms. So it's, you look at the errors, and if some, if it's always, if you

		have a dominance of gender of race, or age or gender, in those errors, then you can, okay, this is you might detect this bias. And before that, there is also the data set representation, when people will look into datasets and see about, okay, if you if, if it's in Facebook and when we just have samples of white males, then , we can already imagine that we are having to have some bias towards white males and not like a black female will, it's likely that she will be misrepresented by this data set, for example. So sometimes it's, it's obvious in the sense, so either in the data set, analyze or in the features that you we can, okay, these might be bias, and then they need to correct this. And sometimes you have a little bit something that are more like maybe hidden to some extent, like location. For example, if you take into consideration like the location for fraud detection mechanism, then which you understand like, okay, it's the same thing, but it's, it might be a little bit more subtle about, okay, just type of frauds coming from third world countries, specific place, like, and it goes again, about the feature importance, and they are analyzed to understand and to make to that will support you in these analyzed to understand how we, whether there is bias, bias or not.
34.	Speaker 2	Thank you. And yeah, the last one is a good example, we had it in in the research too. So I do one last questions in bias. And so based on your experience, what is the most common type of bias? And we're looking more for like technical answer in the way like not gender or race or more like historical data, a representative bias measurement aggregation bias? So what do you think is the most common one? Have you identify? What kind of bias Have you identify at work?
35.	Participant 4	One, let's try to have a broad term for that. But *thinking* I can say that a lot of the a lot of problems that you have some imbalance problems. So I mentioned that anomaly detection, but a lot of like most of the problems will have some imbalance to them. And it's the most when we're plugging in machine learning and artificial intelligence and a lot of these you have a lot of like, the 50-50 data sets and balances datasets. except but in reality, it's very rare that we are we will find so such datasets. So this will lead to bias to a lot of bias this balance this imbalance with each by them. But it's more about understanding the impact of these bias to the business value in at least in the case of industry in research is similar, but you can do the better argument about covering corner cases and different scenarios. But in the industry, you need to understand, so sometimes it's better to just have, like a baseline that you go over a lot of those like, imbalance scenarios, and like major offenders and major case then to cover all of the case, in some kind of it's not, it's not okay. But I guess that they like the representation with data distribution balance in the sense, it's the most common bias.
36.	Speaker 2	Thank you. That's great. And then one more question. And then we go towards three cases. And that's the fun part of the call. So the question is, like, we recognize that diversity is actually a central role in actually detection of bias as opposed like just working individually. And so we're wondering, like, how diverse are your teams? And do you think having diverse teams have a positive effect of on the team bias?
37.	Participant 4	Yeah, yeah, I might seem is very diverse. So we have do you want the specifics about it? Or?
38.	speaker 1	No, just like a general your thoughts of how it affects?
39.	Participant 4	Yeah, I guess the about these represent, like, again, it depends on on the, on the problem, like the benefits of it. Because so for example, if it's a very technical problem, like cell phone, fault detection, it might not be the rest of the team might not be that relevant. But when it when you will, closer to

		some more human oriented problems. So as fraud or advertisements are like these more than text classification, we tend to mention allies, for example, like these are more human oriented, to some extent. So it might it's definitely help them. But again, this is related to the problem itself. So for fraud people from different locations will have different inputs about it. So sometimes, you might, if you have a very gender wise and race, racial wise, they diverse team, but if they are all from the same location, it might not be as good and like, you might still miss a lot of the possibilities, because you have different types of fraud, or what like, let's say, and so it, it correlates with the problem that you that you have. So these like the this type of of bias they are, they might or not, they might or might not affect the performance of the team. So to say.
40.	speaker 1	Thank you.
41.	Participant 4	Yeah, just another comments about that. But yeah, I just have to be clear about it. I think that having diverse teams helps. And but it's just so you just need to be diverse in the it's not some automatic thing like to share if you have a diverse diverse team, it will be automatically better. You it might help and but like generally speaking, having diverse backgrounds helps. So yeah. Thank you.
42.	Speaker 2	Yeah, and that that thing that is virtue especially like, yeah, diversity is not automatically like giving you like freedom of bias. And I think one thing that is great is like education and like having like these backgrounds. So I see you are providing us a lot of information, because as well, you had that into account because you have a study about that, as opposite as somebody that just come into the workplace and work. And so I think that was a great answer. And now we're going to the fun part, we're doing three different cases. And they're quite quick. So I'm just going to share my screen, and you'll let me know if you can see it. And then, here, I just gonna give you a quick explanation. So we have identified that they are deferring steps, what is can happen, and we had divided between like data bias or algorithmic bias. In that bias, we'd say, Okay, could be data collection, or data preparation. And then algorithmic bias could be like in the development in the evaluation, post processing on model deployment, of course, data and algorithmic bias. So the most of the time might go together. And so this is us a general view. And maybe afterwards, we will ask you, like, if you have any feedback about this, but we want you to take these into account when we're providing you with the cases. So if you need to, like read the small definition word, everything is but I imagined, you know, and it's very straightforward. And so we can continue with the case. If See, that's okay. Thank you.
43.	speaker 1	Okay, so this is our first case. I don't know if you prefer to read it to yourself, or if you want me to read it out loud.
44.	Participant 4	But you're no awkward silence. can read it, or you will read together? You can read it.
45.	speaker 1	Yeah. So the first case about is about the man of Asian descent, they wanted to apply for New Zealand passport online. The photo that he uploaded was rejected. And he was presented with an error message that said that his eyes were closed and that the photo did not fulfill the requirements of the system. His eyes were open. And after three attempts with three different photos, he had to reach out to the passport office, the office blamed it on bad lightning and shadows in the eyes. So if we go to the next slide, and oh, no, we can't stay here for ya know, so. So our first question is state, starting from a data

		collection to the model model deployment of the AI system? Where would you say that the highest risk of bias occurring is for this case,
46.	Participant 4	the highest risk would be data collection, but it also might involve data preparation and data evaluation as well.
47.	speaker 1	So if we go to the next slide, we have a piece of additional information that says that there was an under representation of people of Asian descent in the data set used to train the model. Knowing this, would you change your answer in any way? Or would you still go with the answer that you gave us?
48.	Participant 4	You can either collect more data from Asian people, but you can also do over sampling for about the data preparation part and about the model evaluation. If you are using accuracy, for example to to evaluate the model, it might be not the not the best solution in mind because you might have different classes and different ways to under sec to make sure that different race and age standard like these important features they are being represented. So as I said, it's about the tighing the whole thing together like about images need to make right what is the business value? So in this case, it's important that So, again, so you have some bias, like okay, you you have a population population itself. They have different presentations for race, then for gender and for aging, etc. But you We are in like this passport control with no specific place. So in that place you have also this bias. So you need to understand Okay, for this problem. Is it okay? If I just get right the majority like the majority of people? In this case? Of course not. So you need to take that into account and then okay. They they need to look into the corner case and making sure that the all the classes are being well represented etc. Oh, yeah. But it fits for each case was like advertisement company, it might not be a big of a problem if it didn't got those rights, for example. Yes,
49.	speaker 1	thank you, if we go to the next case. So this is about an algorithm called compass that was used in the United States to determine whether or not a defendant is awaiting trial with the reoffend, the decision was based on more than 100 Different factors such as age and criminal history. When analyzing the results, it showed that dark skinned people were more often classified as high risk among the defendants that did not commit new crimes. So for this case, we have the same question starting from data collection to the model deployment of the AI system, where would you say that the higher risk of bias anchoring is?
50.	Participant 4	I would say that data preparation would be the major offender like for this specific case, like the skin color should not be a component, they should be removed altogether. So that should not be a feature.
51.	speaker 1	Yeah. So if we skip to the next slide, we have a some additional facts. The first one is that it could be because of the granularity of data being different. The second one is that they used data on prior eras, from their defendants, families and friends to determine the likelihood of them reoffending. And then the last point is that it's not unusual for minority communities to have a high police presence leading to a higher number of arrests, and drawing the conclusion that the residents of minority communities have a higher number of dangerous people due to having a higher number of arrests is misleading due to factors such as police presence. So what's your opinion, now that you have heard this? Do you want to change your answer? Or would you go with the same answer?
52.	Participant 4	Some, like some of these problem, it's more like a chicken or egg problem. Because even though we remove the race, and again, it's sometimes it's sometimes it's very obvious. Like if you are using race as a input feature to the model, it's, it's very obvious, but sometimes, there's also like other

		features that might infer the race of someone without actually actually saying it. But it's more about the, like, the data is representing reality and reality itself is unbalanced, that is unfair. And it has like the world has racial bias, gender bias, age bias. So it's not like we can just not ignored but it's not like we can fix these in our models. So sometimes it's it's more like a chicken egg problem. And I am glad that I don't work in these cases to be honest, in those problems, because it's way deeper problem and it's sometimes it's basically, we can't automate these we like we don't have a world that benefit enough that we can automate these and avoid justice and avoid unfairness. So my like, we can have like we if you are really careful about data preparation, and looking at the results and doing the feedback loop a lot of times so that we Okay, we are having their preparation. A lot of these are will be related to model evaluation. because that's how you are going to measure the quality of your data collection data preparation. So you can like iterate over these to try to improve and making sure that we're removing part of the bias. But in the end, it's you can't avoid some bias from the real world. Go into this leaking into the module.
53.	speaker 1	Thank you, I noticed that we're past our time. But I was wondering if you have time, we would like to do the last case as well with you.
54.	Participant 4	Yeah, sure. No problem. Yes. Today's a no meeting day for me. Even though I just got a message, you want to have a meeting, but there is no.
55.	speaker 1	Are you sure? It's okay. If you have some words? Oh, sure.
56.	Participant 4	No, sure. I mean, it's, I can have this call later.
57.	speaker 1	Thank you, if the last case. So in 2018, it became known that ads related to steam carriers were less likely to appear for women than men. The ads were run on many big platforms such as Facebook and Instagram. However, women were less exposed to the ads, and they were appearing more frequently to men. So then, once again, the same question starting from data collection to model deployment of the AI system, where would you say the highest risk of bias occurring is?
58.	Participant 4	Yeah, basically, it will be more to Learn on Demand. It's, again, chicken egg problem. In this case, we have less women in STEM careers, and then they will be less interested in on it, and they will click less on that. And then it's, it's so even though Okay, we are just going to remove gender completely from the data collection, like we are, this is not going to be a feature. But because when browsing the site, they have less less interest in part of this not because like, again, it goes to the research question. But let's say that in general, is clear, we can say that. So if nothing is done about it, it will be the same, it will stay the same. And if you just lucky, let's remove gender from the data collection, we would just keep showing the like the to the users the like this bubble behavior, so it will just reinforce whatever they like for fun, like, I don't know, video games, they will show more video game ads. And like books, they show more book ads to make. So it's advertisement definition, it's not supposed to like to put you out of your comfort zone and just introduce you to this completely new stuff. It's more about okay, you are my target audience and I want to reinforce your current behaviors, even though this might be bad, such as fast food and gambling, whatever. So in this case, I guess it's more about the like the the end goal and the problem definition and so, the what if, again, it depends on the business value. So do you want to have more women working on stem and be

		interested on it and having this so you need to take gender into consideration just to make sure that they are going to, like it's an opposite case.
59.	Participant 4	But these three cases are very different than like, in the sense so for example, the first one you need to make sure that everybody is have the representation and everybody like it's important. The second one the model need to be to be more oblivious to the race of the person so you don't want to the much show. Just condemn someone because he is black, or she's black. And in this case, again, it's even if you just ignore the gender it these Information leaking into the model somehow, because it just reinforced the behavior. So if if you want to have more women, women then you need to okay, if it's a women then how we are going to make sure that it will have the same amount of like the probability of having dessert as a man, let's say so or, like, not, not the same, but like higher necessity. Okay, it's balanced right now. So because even if you remove it to, it will not make that much of a difference.
60.	speaker 1	Thank you. So if we present these three new if we change this that, yeah, so the researchers that talked about the previous case presented two possible reasons for this. And the first one is that the algorithm learned the pattern from the consumers, women were less likely to click on the ad, therefore it trained itself that showing it to men would be of higher value. And then we have the second possible reason that is presenting ads to women costs more, and advertisers must pay more for exposure to women. The algorithm is programmed to make the most cost efficient choice and is therefore living women out hearing this, what are your thoughts?
61.	Participant 4	Oh, yeah, I forgot, I forgot to like to make sure to pinpoint the specific but in this case, it's more about the more about our algorithmic bias in to this extent, like the sense because even though, like even if you remove, like, there's not a lot that you can do in data collection, their preparation that you could do to avoid that even though you remove a lot of stuff you would still have similar results or a life that you like, these cheaper problem, like if you over sampled, like the the amounts of like, if you put greater likelihood to the women's samples to have to have stem advertisement, then it would like it would like it would be very hard coded way to fix that. Like, again, it's it's more about the business values and research question like do you want to to? To send targets more to women? Is this something like you want to actively avoid you want to actively improve? Because and then you can we can think about the, the whole process.
62.	speaker 1	Thank you, that was all for us. I don't know, if you have something you want to add or speaker 2, if you want to add something before we finish up?
63.	Speaker 2	Yeah, I just want to say thank you. And I'd really believe that this interview has gave us like so many great insights, I have been taking notes and hopefully this is going to be great for our thesis just want to let you know like everything will be anonymous. So, the name of the company and your name will be deleted from the transcribed transcribing transcription, but if you have any like anything to add maybe to our model or to the whole presentation, just let us know any feedback.
64.	Participant 4	Okay. So, actually one or two questions. So, what is the like, you are looking more into the ethical part of it, like the bias more about like gender bias, age bias or bias general What is your main interest
65.	speaker 1	in the general approach to bias we started off by gender bias, but then we changed it to more general approach.
66.	Participant 4	Okay, good. And how, how many more interviews do you need? Like oh, what is a good number? Because yeah, I mean, I, I can talk to different

		amount of people like I can say that my channel we've like, I guess, almost 100 people or just talking to less.
67.	speaker 1	Wow, we should have gotten in touch with you sooner. I don't know. What do you think, Stephanie? You were our fourth Interview, we still have to to go. But I don't know what you think.
68.	Speaker 2	I mean, we are not looking for that many we're wondering maybe like two more, it's possible. And because of course, we do have a deadline that we need to meet. And we cannot do more 100 interviews, even though we would love to. And so if you know, anybody that is like interested have for next week or this week, we will be very happy to and to use the inner
69.	Participant 4	nurture. Sure. What is? What is the profile? Like? Do you want something like someone younger, more experienced? Like how
70.	speaker 1	we have had all types of people up until now mainly people with a data interests and they work directly with data? We don't have any other preferences than that, I think.
71.	Speaker 2	But I would love to have like an, like more experience is is possible. And because like that, as you say, like yourself, you're where you're coming, how many years experience and we have gotten a lot from this interview. So if you know anybody that has like, that is not fresh out of graduation, and has more experience that will be great.
72.	Participant 4	Okay, okay. Yeah, I can talk to I guess, two or three. Again, it'd be a good number to you that I don't have that many years of experience. But my experience is kind of all over the place. But they had they do have like some like, two, three years, folks that don't like a single problem in, in natural language processing. So we have a lot of good insights to give to you as well. Thank
73.	Speaker 2	you. Yeah. And I send you an advertisement thinking if, I mean, if you need to change anything, I didn't need me to add something. I will, I will do that.
74.	Participant 4	No, yeah, sure. I would just talk to them. So I think that to be around, like, I'm pretty sure that you have them. I mean, I can talk to more, but we feel a little bit more of experience. It's show people in mind that it'd be very good to
75.	Speaker 2	thank you. Yeah. Thank you so much. And we really appreciate it. And we know, you know, the struggle of trying to find people to
76.	Participant 4	share. Yeah, exactly. Enemy. I guess like almost like these two people, they have once completed a master degree a while ago. And other it's a PhD student. I know. I also know the Ag students. I know this guy that has a PhD as well. So like, we know that. Yeah, I know. It's, it's tough. And like you are doing, is it a master's degree or PhD? Masters masters? Oh, yeah. Because you have these like, you can do like these in pairs, right? Yeah. Yeah. Yeah. That's a good thing. Because I used to say that my PhD was the hardest thing I ever did. Love that. Place was, was close to one where I did my PhD. So yeah, it will get easier. Hopefully.
77.	speaker 1	Yeah, actually, I started this master program thinking that I wanted a PhD, but I changed my mind.
78.	Participant 4	But it's good. It's good. Like, it's it's fun. And funny, it's a pity that you are not walking with us at Sage because we do have a lot of master's studies from loans kGA and a lot of places. So yeah, we know the struggle. Yeah,
79.	Speaker 2	then like, the thing is, like we this is one year master's degree. So it's very, like condense and we had like use a few months, not even not even like six months to actually work on the research. So it's very like we couldn't pay or win on the Frank company. But at the same time, it's it's a good because it also we get the freedom to choose whichever topic we want it. And yeah,

		we felt that this was very interesting. And there is like a lot of information online, but at the same time, everything is telling like, oh, artificial intelligence is great. But then we went to like Then we go into like the pirates like Yeah, but what if is not so great? And that's like, I think that will be great when we talk in our when we present it. Because people is always expecting like the benefits of it.
80.	speaker 1	Yeah, sure. It's
81.	Participant 4	I mean, I'm pretty sure that you do. It's very, very enriching experience to have these videos and talk with different people. And yeah. And I think that the Mike's my teammates will be able to give you good info, as well. But sorry, sorry, if I have a little bit of sketch on sometimes about artificial intelligence and machine learning. I'm a little bit grumpy maybe. I'm not that enthusiastic about Yeah, machine learning will solve everything will. It's the best thing ever.
82.	speaker 1	Yeah, but that's what we're looking for. We want the other side of it.
83.	Participant 4	Yeah, yeah. Great. Yeah. Okay. I will talk to them. And I guess I will, I guess I will ask them the permission so that you can send you can talk to them. Instead of like this sending emails they will give, I will give you their emails. And then you can arrange the interviews with them. Okay.
84.	Speaker 2	Is it amazing? Thank you so much. I hope you have a nice day, and then we'll send you the transcript when we have it. Yeah, okay. Thank you. Bye bye.

Appendix 12: Transcription - Participant 5

1.	Speaker 2	Thesis is about bias in the context of artificial intelligence systems. So the focus will be to tackle this from a data point of view. So we want to find out the effects of data on the outcomes or the artificial intelligence systems. And we want to actually identify where in the process are the bias more likely to happen? And so of course, we know that there there might be a lot of influence and a lot of factors. But then we want to pinpoint a specific process a specific part of the process. Yes,
2.	speaker 1	so the interview should take about an hour, and you have the right to stop it whenever you feel like it. Or if there are certain questions you don't want to answer, that's okay to just let us know. And as Stephanie said, and as you're seeing the call is being recorded, but it will only be used for this study, and later, we will delete it. We will also send out the transcription that you can read just to make sure that everything is right. And there you have the opportunity to change or remove anything you'd like. Okay, sounds good. So let's start with some warm up questions. Could you give us a quick introduction of yourself?
3.	Participant 5	Yeah, so I'm X. And I did my undergraduate in mathematics in the X. And then I worked for two years as a financial analyst. So I have a bit of background in that. And then I got my master's in artificial intelligence. And I did my thesis in explainable reinforcement learning with causal models. So somewhat, not related to bias, but with explainability. So like that power of AI, and now I'm working at X as a data scientist.
4.	speaker 1	Great, thank you. Could you tell us something about your experience with data and artificial intelligence?
5.	Participant 5	Yeah, so that mostly started our work while I was doing my master's. So that was all based on AI, learning the different algorithms what to use them for, we had a class for AI in society. So there, we learn more about the biases. And the GDPR, for example, kind of related to biases and the rules to try to avoid them. And here at work, not all of my projects are related to AI or data directly. But we're doing a couple of projects in which, in which we're building models that identify damages on pictures. So we're not dealing with personal data, but still managing data and building AI models with it.
6.	speaker 1	Thank you. That's interesting.
7.	Speaker 2	Yeah, I, I kind of start with, with the first topic, which is AI. And we have two questions here. And we are going to give you a definition, how we feel that we define AI. We know that it might be everybody has a different definition, it's actually hard to define. And so our definition is more towards society. So this is the definition and then you tell me if you want to change something. So an AI is the ability of a system to identify, interpret, make inference, and learn from that to achieve terminate organizational and societal goals. So do you agree with this definition? Will you add something? Or change or remove?

8.	Participant 5	I'd say I mostly agree with the definition. That's what we would want AI to do. Yeah. Ideally **smiles** Yeah.
9.	Speaker 2	And if I were to ask you, how would you define it in your own words?
10.	Participant 5	I mean, I would say, pretty similar the ability of systems, mostly computers, to achieve tasks that humans usually do in order to improve our lives. And automating processes.
11.	Speaker 2	Yes, that's a very important power. And so the second question is like, related to that we have read points out that there is a need for addressing the dark side of artificial intelligence, for example, as we're addressing the bias in the systems at the bias. So is this something that is actually talking in your organization and with your workmates?
12.	Participant 5	I think within the job for the six months I've been here, it's mostly talked about within the context of the GDPR. Like it's more about we have to be in line with the GDPR. Yeah, that's we don't it's not talked about in the context of like making sure that the models do what we want is usually just based on the GDPR.
13.	Speaker 2	Okay, so we'll just say then is because of the law, then you meet the requirements of the law, but otherwise, it wouldn't be talked that much.
14.	Participant 5	I mean, personally, my personal opinion is not that, but I would say, in my experience so far, yes.
15.	Speaker 2	And what is your personal opinion?
16.	Participant 5	I'd say personal opinion, it should be thought about more in what are the effects of the models that we create, not only in the context of the GDPR, but more in the context of, from a personal perspective, the effects you will have in society
17.	speaker 1	Yeah. So now we have some data related questions. Maybe some of them are not, are things that you haven't worked with, but that's okay. You can just say that or if you have other things to add. But the first one is, when collecting data, do you have any specific requirements that you follow, or other aspects you try to consider?
18.	Participant 5	I think, within work, the most important aspect we consider is making sure that the data we are collecting is relevant for the use case. So for us, I mean, not related to biases, but for example, in the project, where we are collecting car pictures, to then be able to identify the damage damages, we have to make sure that the cars are gonna be like the cars that are more present in the country where we're gonna use the model, because our client is outside of the European Union. So maybe the most common cars are not the same. I mean, cars are cars, but it's important to train the model with like more data coming from the specific country.
19.	speaker 1	Yeah. That brings me to my next question, how do you make sure that the data that you use is of high quality?
20.	Participant 5	So I think, yeah, that one of the points is making sure that it's representative of the group where you're where you're, where you're going to use it on. Then I would say, of course, during the pre processing step, making sure there is no faulty data and not missing points? Or if there are some missing data, have a process on how to manage that, like, are you going to delete those data points? Are you gonna use an average? Are you gonna try to find the right value? So? Yeah.

21.	speaker 1	Thank you. Um, so could you describe the process from when you get access to the data that you're going to work with until you use it for your models?
22.	Participant 5	Yeah, so following the example of this model are building we get the data with from the client? Yeah, we just have to do the pre processing, make sure all the points are there. I mean, it's a small project. So honestly, it's like a short process. We just get the data, make sure it works, make sure it works, do some pre processing, and then start building our model.
23.	speaker 1	Okay, so the client gives you the data, and then you do the quality checks and stuff like that.
24.	Participant 5	Yeah.
25.	speaker 1	Would you say that you could ensure that there is no risk for bias in the data that you work with, or the algorithms or systems that you create?
26.	Participant 5	In general, there are some checks you could do. And if you, for example, identify biases, there are some algorithms that you can apply to try to mitigate the biases. But I would say you can never be 100% sure that there is no bias at all.
27.	speaker 1	Yeah, so anyone who says that is wrong?
28.	Participant 5	I mean, there's a way to, like ensure that it's pretty unbiased. But yeah, I would say, I guess someone that says this is for sure. 100% unbiased, I could be wrong.
29.	speaker 1	Yeah, no, we agree with. Okay. So our last question regarding data is that when you work with demographic information, do you work in a more precautious way? Are there stuff you take into account?
30.	Participant 5	Yes, I would say when you're working with demographic data is when you should be the most careful as that's usually where the biases are at. So, when working with that, I would say it would be more important to do more pre processing steps in which you take when the different like qualities of concern if there are any biases and if you identify some biases, maybe Apply some weighting algorithms, things like that to try to get rid of the biases as much as possible.
31.	speaker 1	Okay, so that means that when working with demographic information, there is a larger focus on the pre processing part.
32.	Participant 5	Yeah. Yes.
33.	Speaker 2	And I have a follow up question in the data part, actually. And I understand that you say that the client gives you the data, and then you process it? Do you give some requirements to the client? Or do they give you like the raw data, and then that's where you like, pick up the specific ones that you want to use.
34.	Participant 5	So for us, we actually give them like, requirements that try to avoid any personal data being on the pictures. So since we're dealing with car pictures, we ask them that no people are present on the pictures. And we ask them that they try not to get the license plate, since that could be considered

		personal information. And in the case that we see there is like license plate on the pictures, we try to we blur them.
35.	Speaker 2	Yeah, perfect. Thank you. And I can continue with the bias part. And so how would you work in order to detect bias? Like, do you follow any protocols? Or how does your organization can work towards the denting bias?
36.	Participant 5	I haven't been involved in that process. So I can't answer from like a work perspective. But I would say if I get a data set that I can see has demographic information could have bias information, I would run some of the algorithms that detect these like, or like run to get the key metrics of demographic parity, and all those to make sure like there are no biases. And if there is, of course, I wouldn't want my model to be biased. So I will try to fix that.
37.	Speaker 2	So I know that you don't work with it. But is your company actually able to identify the bias?
38.	Participant 5	I would say, within the scope of the detecting with the regular algorithms, the ones for example, in the AI 360 library. I mean that, yes, if it's, like bias within the data in a very hidden way that it's not recognizable within the regular algorithms. I don't know. I would say probably, if we're dealing, I mean, at the company, we work out with banks, and financial institutions. So I would say, we're probably dealing with a lot of personal data. So I'm sure there are, like checks in place. As I mean, we're in Europe, we will have access to personal data. So I'm sure there are checks to make sure everything is done the right way.
39.	Speaker 2	Perfect, then, maybe you might not be sure about this one. But just in case, I can just throw it in there. **participant smiles** And based on your previous experience in education, or at work, what would you say would be like the most common type of bias? Like, here we we are looking for a question more technical in the sense. Is it like a historical data representative? Sampling that sampling bias or measuring bias?
40.	Participant 5	Or so this from my opinion, but opinion, it would say probably the biggest issues in historical data. And the fact that the historical data is bias itself. So then it's it's more like we were, you could say, we were biased in the past. So now all the data is biased.
41.	Speaker 2	Yeah, that's basically where we actually trying to confirm as well, that it's a bigger issue that lies in society rather than the system.
42.	Participant 5	Exactly. It's not that the specific data is biased is the fact that if we get historical data, like, yeah, men, what had the tech jobs, for example The algorithm will see that.
43.	Speaker 2	Yeah, yeah. And I can continue with like one, we have one question on Team diversity. So we recognize that team diversity plays a central role in detecting bias as opposite of working individually. And so we want to ask you like, how diverse are your teams? And do you consider that having diverse teams can have have a positive effect on the death in bias?
44.	Participant 5	I would say my team from a gender perspective, it's actually pretty balanced. I don't know. 5050. But close to it, probably. And in terms of, for example, nationalities. It's pretty diverse. But within probably like the European Union, although we have other, we also have people from other countries. But I would say that's the Yeah, usually within the European Union, very diverse could be more diverse in terms of worldwide nationalities. But I would say it is very important to have people from different

		perspectives, to be able, because if you know something, that it's usefully like an issue, you will look for it in the data, if it's something you've never think about, maybe you'll miss it, because you've never thought about it. So different perspectives will definitely help.
45.	speaker 1	Yeah, the literature taught us that as well. Yeah.
46.	Speaker 2	It's good. We can confirm everything that we read.
47.	Participant 5	so opinions, it's not confirmed by my work. But yeah. *laughts*
48.	Speaker 2	But no, I think that's great. And then now we have the fun part that we're going to present you like three different cases where bias had happened in real life? And maybe we can share screen
49.	Speaker 2	No. So here we have, we have created this model, where we have identified that , there is data bias and algorithmic bias. And data bias could happen in the data collection, the data preparation, and alogirithmic bias could happen in the model development, the evaluation, the post processing, or the deployment. So then we can have on this, if you need to read what each one it's about, that we had, like, there the numbers, but then it's very straightforward. So you let us know and then we can share.
50.	speaker 1	Yeah. So this is the first case. I don't know if you prefer to read it to yourself, or if you want me to read it out loud.
51.	Participant 5	I could read it. Yes.
52.	speaker 1	Okay, yeah. So we have a question, then, starting from data collection to the model deployment of the AI system, where would you say is the highest risk of bias occurring, this bias, like for this case?
53.	Participant 5	data collection
54.	speaker 1	Okay. So if we go to the next slide, we have some additional information. And knowing this, would you change your answer? Or do you still think that data collection is the reason?
55.		This confirms that Yeah.
56.	speaker 1	So let's go to the second case.
57.	Speaker 2	I was wondering if you actually can use explore a little bit more the answer you say? Like, why do you think is that a collection?
58.	Participant 5	Yeah, I would say because since there was an under representation of Asian people on the dataset, the algorithm learned that the eyes open, were bigger, because everyone that had eyes opens on the pictures they collected, the eyes were bigger. So then when it sees like, I guess, really small eyes, it's not able to recognize that that's open, because you know, they labeled pictures it's been trained on that wasn't an open eye.
59.	speaker 1	Thank you. So this is the second case. And the question is the same. Starting from data collection to model the deployment of the AI system where would you say is the highest risk of bias occurring?

60.	Participant 5	I mean, I would still say still say data collection. But probably not exactly for the same reasons as before, or not like as straightforward as before. But still data collection? Or **thinks**? Yeah, data collection.
61.	speaker 1	Yeah, you want to share your thoughts with us?
62.	Participant 5	I mean, it's a still? Probably not. This is probably not an issue of under representation of people of dark skin. It's probably more an issue of historical. Yeah, the history of the history of darker skinned people in the United States, and maybe historically, in the past, they've committed more, or they've committed crimes more often. So I mean, since they said that's on the data, it's data collection issue, but that shouldn't be used against them for the future. Yeah, difficult one.
63.	speaker 1	Additional facts to this one as well. So reading this, we would like to know your thoughts about this?
64.	Participant 5	They would say to still data collection, and it's, yeah, it's more the issue of Yeah, historically, there are more police there. So they get arrested more, it seems like.
65.	speaker 1	Thank you. And then we have one last case. Yeah, this one? Same question here as well.
66.	Participant 5	I would probably still say data collection and maybe also model evaluation. As Yeah, same thing, this is probably because the data they in the data, they used STEM careers, or men had STEM careers in a higher percentage than women. So then the algorithm thinks that the men are more likely to click on it. So it's an issue of the data they were collecting. And I guess also the data preparation as they could. But this could apply to the other examples to data preparation, because they could have pre processed this and noticed this, and then fixed it before doing the model. So it's a bit of data collection, data preparation, and then model evaluation in the sense that they could have checked this before deploying the model.
67.	speaker 1	thank you. And this case also has some additional facts. So we would like to hear your opinion regarding these as well.
68.	Participant 5	Okay, then I guess this actually, I would maybe change my answer. Yeah. Probably model development in that case, since the the model was actually built to work that way.
69.	Participant 5	if they were asked, but actually, I will say, i dont know if this is correct to say, but if they were actually less likely to click on it, then I mean, I don't know. model development. Okay. Yeah. Because it was trained to work that way. There weren't any checks to make sure the algorithm wasn't doing this. Yeah.
70.	speaker 1	I mean, these are just like potential reasons. We don't know the right answer either. The author's who studied this case, just listed these as potential reasons. So that's why we're interested in hearing Yeah, yeah. But that's all for us., do you have anything you want to add?
71.	Speaker 2	Um, no, I just want to say that, yeah, it's very complicated to actually try to pinpoint specifics, because especially there is like, so many factors happening. And sometimes it's more than one. So So basically, data collection might be something that model evaluation should see. **participant agrees** Yeah, but then people, like, just take that for granted. And then they release systems that are bias. But then with this, we just want to have like, different kinds of opinions. And I think it's great that you got to

		actually give us some insight, mostly on this on the side that you're, you're saying that you guys had these specific things to find the that find the bias, and, like, mitigating Explainable AI, that's the one that I that I was like, looking at, right. **participant agrees*. And, yeah, that's also in the literature that we have read that Explainable AI is something that is like getting more popular. And and with these, we just want to say that all the all the backgrounds are necessary, and to basically try to come up with a with a good thesis.
72.	Participant 5	Yeah, no, but it's a very, very interesting thesis, I would say.
73.	speaker 1	Thank you. So it's good to I'm sure. Yeah. Thank you.
74.	Speaker 2	Do you have any like, advice or any like last topic that you want to talk about? Or maybe you want to add something?
75.	Participant 5	I know I would say I thank you very much for this. It was very interesting to see like your thesis, how you're working on it, the topic. I really liked the topic. So it's, it was very interesting.
76.	speaker 1	Thank you for participating. Thank you so much

Appendix 13: Transcription - Participant 6

1.	speaker 1	We will send out, we will send out the transcription so that you can read through it if you want to make changes or add anything. And other than that you also have the right to cancel the interview whenever you want or skip certain questions. So that's okay, too. So with that being said, we could start with the first question. Could you give us a quick introduction about yourself?
2.	Participant 6	Sure. Um, let me see. My title changes. So right now and machine learning engineer. I've been working with data, I don't know, three years and a half round. And in way, I'm data scientist, now also a tutor in the biggest university in Brazil. And I am professor in digital school. So I used to teach the machine learning.
3.	speaker 1	Oh, great.
4.	Speaker 2	Yeah, I think it's great that you have a like the experience and in their, in their business, but also, you have the academic part. And this is something that we are really into it in this research.
5.	Participant 6	*laughts* Nice. A good match Actually. Oh, I used to think about, I would like to see, I was thinking about to be a professor. But I didn't fully like the academic. Of course, I liked the academy, because I'm a PhD candidate right now. But I don't know the businesses is completely different. So you can see the use for this teories
6.	speaker 1	Could you tell us a little about your experience working with data and artificial intelligence?
7.	Participant 6	Okay, I started around two years ago, with work with data work as a data analyst. So mainly doing dashboards and reports and things like that. And after one year working data, as a data scientist, I am in a job as a data scientist that I've been working with. In right now, working with machine learning. And statistic is, as a data analyst is more reporting dashboards, of course, they work with statistics, but it's, it's how statistically is a gross way. So it's a medium average. So it's not so deep as machine learning is.
8.	speaker 1	So the role you have today is more technical. **participant agrees**. Thank you, you want to continue?
9.	Speaker 2	Yes. So we going with the artificial intelligence part. And I'm gonna give you a definition of how we define it, which is a bit more social. And then with this definition, you will let me know if you agree to it, or do you want to change or remove something, and I'm gonna read it, but I'm also gonna copy it. So like that you read it, because I know sometimes is hard to understand. So, we define AI as the ability of a system to identify, interpret, make inference and learn from data to achieve or terminate organizational and societal goals. So what do you think about definition this definition?
10.	Participant 6	Okay. I disagree. In somehow I agree that is more machine learning definition. Just let me explain a little bit about AI. It's more broad. So AI doesn't need data, for example, event thinking as automation, I have an assistant that turn off or turn or turn on the lights in my house. They doesn't need data to train or to do anything. They just need a light sensor to detect to detect if it has, I don't know sunlight outside to turn off or turn on my my

		lights inside of the home. But when we think in machine learning, they they it's a It's a constraint. So it needs data, to learn to try to identify something. So I saw AI as a huge area. But we can go deep in Data Mining and Machine Learning, or either on deep learning and so on. So, I agree but disagree. I guess this definition is too specific for AI. For example, I guess a Russel is the name? It's a famous book, I don't remember the title
11.	speaker 1	We actually have that book as well.
12.	Participant 6	He defined AI as something, an AI intelligence demonstrated by machines. So it's, you can see by this definition, it's it's broad. Everything that is, as a simple intelligence could be an AI. So if my problem is a light that turned on/off automation it's simple if and else structure so but it has an intelligence on it, because I developed a intelligence that should turn on or turn off the system. So I don't know. I will. I agree. But I disagree. I most agree with this definition, I guess this is the machine learning definition. Ai, it's more generic.
13.	speaker 1	Yeah, we're familiar with the Russell and Norway definition as well. Would you say that you agree more with that one?
14.	Participant 6	Yeah, I like more Rusell definition.
15.	Speaker 2	They, I think it's a very, it's a very, it's a department like it's a whole study, and it's very hard to define and things that could be aI before, they are not AI anymore. **participant agrees** And so yeah, this is like, basically quite general, like, we just find one that we like, that wasn't that technical. And they were like, Okay, let's try to see if everybody's agree or disagree. And these are great insights that we can use, always use in our research. So to continue, and the literature that we actually read and say that there is actually a need to address the dark side of artificial intelligence. So with these we mean like AI systems, adding bias, so we wondering, is this something that is talked about at your workplace with your colleagues? Or in general, would you like with your friends?
16.	Participant 6	Good Question? Actually, I, I didn't work so close with ethnicity AI. So, we always has, we are concerned about so we, of course, it's important to think about, but I i lie if I say I always thinking about that. So, I not **uhmm** but we in some times is really important to take a look at that. So we can see a lot of examples, for example, face recognition that doesn't recognize black people or whatever. So, I guess I never work with this kind of problem. So, with a problem, let let define that with a problem that has high viability of social people for example. I, I worked with, medical person. So basically, basically, the way, I used to work with text just to to explain that. So the way of the doctors write messages on WhatsApp, your SMS, it's pretty similar. But of course, if you're working with the huge population of the world, we we are more, we are less, less formal. We are really informal typing messages . But most of my projects are focus in some area. So somehow it's good to be biased by that area. So if I develop if I'm working with the sentiment analysis, so sentimental analysis on. I don't know. In Amazon reviews is completely different in sentiment analysis for doctors messages , so I guess bias, it could help us in somehow. It's strange

		to say that. But you can, you can, we can specify some some applications, according to bias. Do you mean bias by social bias? I'm right. Okay.
17.	Speaker 2	So more like discrimination.
18.	speaker 1	No, I get what you're saying. And we also had another interview, who pointed this out that like, for some cases, removing gender could actually make it more unfair, because that special data point could contribute to a better analysis and understand the results better. So that's actually a valid point that we have learned during the interviews.
19.	Speaker 2	Yeah. And another one that I read about it, it was that you're advertising a job that maybe is not good for older people. So of course, you are gonna add bias and that because you know that you don't want older people to actually break a bomb or something. So it's, it's weird, but then he actually helps a little bit. But as long as he's not discriminating somebody, we're fine with it. But just to confirm, I know you're not working on it at the moment, but at the in the workplace is not something that is being talked about, in general.
20.	Participant 6	Not a lot. No, we are we are more. No, no social bias, not not.
21.	speaker 1	And it's not a topic you're talking about while teaching either.
22.	Participant 6	No, I am more focus on machine learning stuff. So more math, and statistics. Yeah, of course, we hear a little bit about so. But I didn't go further to truly understand the area. It's a recent talk, as well.
23.	Speaker 2	Yeah, exactly
24.	Participant 6	pretty recent.
25.	speaker 1	So then we have some data related questions that were that we could ask. So the first one is when collecting data, do you have any specific requirements that you follow or any other aspects to try to consider when choosing the data to work with?
26.	Participant 6	Interesting! Well, let me think, the most recent way because we have different kinds of projects, but they are the perfect project for any machine learning engineer, since we receive a bunch of data
27.	speaker 1	from the clients or you collect it yourself?
28.	Participant 6	usually, we yes or no. Let, as I as I said, we have different kinds of projects, but let me think in an outside projects, so for some specific client, no, most of the of our clients didn't send us messages usually, usually we have some system for them, let me think in a chatbot something that is easy to think about. So, we do have access to data. So, of course with requirements and consent about the client, so that they provide us access to data, but the system ISAR, so is RS or we can read the data according to to the rights of the client. And after get this data, so, I manually get this data from your system. We try to to understand what the clients need. Let me think an example I want to classify messages gender age five for user wants to buy a product or or they want to, I don't know, talk with the customer service. So I just want I it's a binary problem, I want to know if he wants to buy a

		product or he has to talk if the customer says, basically, we got the data and we I said we as a data scientists instructor that this data and send it to the customer so hey customer, this data saying that we found in your your system, and what do you think about do you have another examples or fleet. Do you have more more examples in another system or i dont know some kind of mess that a he usually receives his own instagram or in Facebook in the other platforms that you want to provide us to, to enrich our data. So we have a bunch of manual back steps that we are doing to send to the client client send to get, and then we are you we are going to we, we analyze the data over and over again sometimes. And what what important character is that, that we have to, we are usually trying to look for, it's variety. So I want to have a variety of kinds of data. So different ways to try to reach customer service or different ways to try to bind things through message. So I guess that's it. Did I answer the question or?
29.	speaker 1	No, I think you did
30.	Participant 6	it is data related. You mean this?
31.	speaker 1	Yeah.
32.	Participant 6	Okay. So usually, we, we go over and over trying to send a client receive and send them again, until reach a good data's data sets with different different ways to typing text for each of the classes,
33.	speaker 1	and by good, you mean, like it's a data set that is good for the customer? Or like it's a data set that is of high quality? How would you define a good data set?
34.	Participant 6	Oh tough question, really tough! Yes, we, at least me, I focus on the variety of the data. So usually, I looking for different words, use it to express something similar. So what's the different ways that you want to that you express yourself to buy things service, I want to buy something I interesting in I dadada in this product? So for me, it's, it's not it's we cannot quantify that. It's a mark quality. So for me, quality's different words are different two ways to express the same thing. So and sometimes it's difficult, for example, to get a big data set, usually, it's better has a small data set with your high varity of the data, instead of to have a huge data set with much more equals messages. So everything is saying the same, you know, I want a blue shirt, I want a blue t shirt, I want a black shirt, it doesn't mean so much because they are similar. So in this case, it's a bar data set, I guess.
35.	speaker 1	Yeah, cause this brings me to my next question, how do you ensure that the data you use is of high quality?
36.	Participant 6	Um, I guess the, the best way to do that is putting on production. So usually, when everything that we doing in laboratory, it's how it's not good enough. But when we the when I put something on production, so I finished it that chatbot that you try to classify customer service or buying texts, we really found the borderlines. So where do the algorithms go wrong? What are the missing classes or so on? So I saw I mean, that machine alone is not one thing that you do once and it's over, we have to iterate over and over and from that definition, I guess is tended to infinity, but

		of course, we cannot do it to infinity. Sometimes we we we do a few iterations. So three Oh.
37.	speaker 1	So it's in process, where
38.	Participant 6	basically we are we are evaluated, we are manually evaluating the answers from the users. So I are not thinking the client as a company, but I think in their final users, the clients of our client so And usually we have a data scientists. product guy, so someone who developed the Chatbot in one one people, one person from the, from the organization or for the company. So these three district guys try to, to read the data. See the the answers from the from the final customer? And we have we do have some metrix is to see how is our missing labor? To see we have? How is the name? It's something? I don't know, it's like I don't know, answer. So when the Chatbot gets get lost, so they does know if it is a customer or is a buying? They said, I don't know what you're talking about. So if they, we have a high, I don't know, answers from the chatbot so you know that you're doing something wrong haha. And we had to fix that. So usually, usually it's it is a lot of manual evaluations. So read messages for is that to work with text, we have to read messages.
39.	speaker 1	Yeah. And our next question, I think you've touched upon it. But could you describe the process from you get access to the data you're working with until it's used as input for the algorithms or models you're using?
40.	Participant 6	Um, we have different ways to do that. So when I'm working for a specific customers, usually we have to create a contract to explain what are the data that you want to access. And the customer has ability to see, the actual message that we are reading. This is vary from country to country. So for example, in Europe, we have GDPR law to, to, to person so I cannot read for example, telephone number, or sometimes even Name. So we have channel name design data, before the data scientists actually read the data set. But in other countries, I don't know like Mexico, we actually can read the personal data, because they have different gpdr law in this country. So this depends. And it usually when you're working with a customer, it's more tough, because they have different contracts and laws and everything to do. And sometimes it's really restrict, for example, I'm leaving Brazil. So but I work for a global team. But I cannot access Europe data, because it is it's right on GDPR its definition only people who live in Europe can access Europe data. So I cannot direct access the data. I have access indirectly through anonymize data. So I can access data if it doesn't appear any person on name email sometimes website depends of the kind of that website if yes, for example, some authentication code, we have to anonymous IDs authentication code, so, I cannot access the data direct and even European guys, they have some restrictions.
41.	speaker 1	Yeah
42.	Participant 6	So, for customers, we have a lot of depth analysis regarding the country regard the kind of data and usually we have to do some contract to to have this access to the even there were as I said, We for example, one of our products developed chatbots we develop the the Chatbot. Of course, the data is passing through our chatbot but I cannot access it without this contract. We needed this contract to to access data. But when we are thinking

		internal, so some products are for ourselves. So for example, I developed a system to detect if one of our products sent SMS, so a developer assistant to monitor every operator from every country for every place. And so usually I don't know I are 20,000 ehh kind of ways that you send the SMS through the Word. And I create a machine learning algorithm to detect when we are dropping off the conversion rate, the success rate. So how many messages is sending through ever, ever? We call it broker. Every broker, it's our way to send message. It's a combination of country operator, and so on. And so to work with this data, it's it is because it's an internal data so I can fully explore the data. I have the access, and I can get more data to combine with this data. So it's easy, but to work with customer everything, it could be, it has to be under contract. So we need to elaborate a contract and follow the contract rules.
43.	speaker 1	Yeah. Yes, and will have a yes or no question now! Would you say that you can ensure that there is no bias in the systems that you create? Not you personally, but like at the company?
44.	Participant 6	At company? It's tough. It's real tough. But but in general words, I said, No, I cannot ensure that. But it's really tough, because I don't know 3000 People work in that scene. So I don't know what everyone's doing.
45.	speaker 1	That's the most common that's, I think that's the only answer we have heard until now, so yeah. And then we have one last question for this section. Do you work in a more precautious way, when you're dealing with data that contains demographic information? Or like are there any requirements you follow on how to work with it?
46.	Participant 6	Of course, oh, this actually, this is a tough problem for machine learning. It's because I'm, as I mentioned, most of my projects I relate to text in, we can clearly see difference over, for example, the country or even in the state of a country, we have different ways to express ourselves, we have John, here's your analogies. So we can see that, for example, United States, that's a big country. Usually, they the Americans have different ways to say the same thing. So that's difficult in some times, right now for example, I in developer system to worldwide so everyone comes every world, you use this, this system. And it's it's different than the way that we send message through in Brazil, it's completely different than way that we send message in French or in Sweden, or whatever. So usually what you do, we are customer oriented. So if we are most of our customers are Brazilian, we have a big data set for Brazil. And over the time, we are adding more customers to this system. So right now, for example, we add a French company in our system. And we are not we are now collect data for French people to try to fine tune the machinery models for French as well. So it's really important and we can clearly see the performance in going down in we try to generalize the problem we have to work with data for all countries. And one important thing as well is the balance of data. So I cannot have a lot of data from Brazil only and a few samples for from other countries just doesn't make a good model for this. So it's important to have a lot of data from different ways. different perspectives, different countries, it's really important. And it's tough. It's really complicated. collect this data. Or as as you question about sometime is we we do receive data from the customer but usually they send it I don't know 100 is the biggest data set that customer

		provide to us the 100, and 100 for any machine learning it's nothing. It's it's it's almost nothing. So yes, we need and it's really important, but it's tough to collect every little
47.	Speaker 1	Thank you, Speaker 2, you want to take over?
48.	Speaker 2	Yes. Now it's, we're going to focus on the bias part. And so how do you work in order to detect bias? You or people in your organization?
49.	Participant 6	***sighs*** I usually work, I usually try to identify bias, but I guess I'm doing a how analysis, for example, I can I can split my data into different countries, but I don't know, for example, how many, men's or women's send that message is so, but I don't know, the specific nationalities, because we do have some difference between states in the same country. But, for example, thinking in Brazil, the most popular state in Brazil is San Paulo. So, of course, most of the simple measures that we collect is from San Paulo. So we already have some problem from regarding the How can I say that the true the yes, the the people who is using our system, so I don't have examples or samples from every, every country, I have some some bias data regarding specific place that most that are the most popular users. And we do have the same for for countries from the country view. So right now, we have more, for example, in more customers in United States. So of course, we are going to collect more data for the United States. But it's important to collect data from United States from any country in the European country, in South America, and so it's important, but it's difficult. For now, I only doing how visualizations to see how how it's working. It's complicated. **laughs**
50.	Speaker 2	Yeah, and I also feel like your product is very like word embeddings. And so you're only using words. And so what I remember that we read about bias in words more towards like, for example, word associations. So if I say in English, like doctor, and it will appear mostly doctor to be related to men, rather than to female, a nurse, great female, rather than to men. So that's how I've seen that maybe aaa it can be related to your work, and this kind of bias. And I know, for example, I speak Spanish and I know, Portuguese is like you, we do have the feminine and masculine. So we know we're talking about our female doctor, because we put the A at the end. And so I will say like in English, it'a b bit more confusing. Or maybe take these into account, if you like, maybe when I asking you like the next questions, so like that, it's, it's a little bit more like, okay, because it's hard to understand bias in words.
51.	Participant 6	*** laughing***
52.	Speaker 2	And I know I mean, words like you don't even need do you even need like to know people's like genders or people's ethnicity? When you're using the chatbot
53.	Participant 6	Usually, we don't, we don't, we don't need to see that. So and thinking a little bit about our systems. And usually we do we have a process in wordering process that is called tokenization problem. And in tokenization, we usually remove the suffix and prefix so even in Portuguese or Spanish or any language that they have gender will remove the gender part. So we are not concerned about we don't we don't need to usually we don't need to specify if it's you a man or woman or are you gonna keyed so but of course

		we have concerned about but it's more, I guess, for it's more country. So we are more for example, in English in United States completely different english from Brazil or for Africa. So this is our concern the the way that people express ourselves, but not on but not the personnel level. We are not concerned about the gender or the culture or no, it's more related to culture than the gender, I guess. Because the culture, it's region basic.
54.	Speaker 2	Yeah. And I remember that you you said that a doctor will express themselves, many medical staff different to how I express myself. And even academic people will express themselves in a different way. And with more fancy words than the ones that I will use
55.	Participant 6	indeed ***laughing*** i hate that.
56.	Speaker 2	I know. Yeah, to continue and, then let's say, Will you say, eh your organization is able to identify bias or if bias, what is the most common? And here I looking for a answer more technical? So maybe you like this one. So we're thinking like, historical bias eh sampling bias, measurement bias or bias happening in the algorithm? So where do you think is the most common type of bias happening?
57.	Participant 6	Umm, Yeah, I guess the data set bias, oh, it's tough. Can you repeat the kind of bias that you mentioned?
58.	Speaker 2	Yeah, so we have historical bias that is more like about the data, which is like a if there is more male happening, taking, taking technology jobs. And even though the whole data set is complete, there is still like a lot of male on the data set, then representation. sampling bias is, of course, you have collected data, but you cannot collect that enough data from any specific culture. Then as well, measuring bias is more like in the preparation of the data. And so it's in the data preparation, where they're, I feel that, let me, let me think about that one, because I don't have it in the top of my head. That one is more of, like, when, when you're preparing, and then you're cleaning the data, and you kinda over clean it. And then you say, okay, there is a specific measurement that we didn't count for. And then there is like a few other ones, that we're not focusing that much. But those are like what's happening in the algorithmic, maybe, and how you deploy your systems, you created a system, and to check on something, and then you use it for something else. So that creates a bias. And so those are kind of like a few other ones that we have identified.
59.	Participant 6	I guess, for most of the projects that have been participating is really related to the collecting parts. So it's, it's simpler, for me try to identify bias in the algorithm part. So after after the data, so I have the data, but I can want to check if there are which means, for example, biases to one specific class or bias to analyze some pirate specific parts of the text instead of try to read the all the texts. I can check that. But regarding the data, it's it's the most complicated thing for us because sometimes, as I said, I can collect Brazil data, but most of the person who use my data is it's it's living in one specific country in one specific state. So the representation of this data in somehow it's not good enough, because I only have Brazilian data from this state. But there are other states that people have different ways to express themselves and in this case, we are going to be wrong we are doing wrong analysis because I cannot use the same predictor classifier that are

		developed for one stage to predicting another because they used to write in different ways. I don't know if it's difficult to say that but we have we do have a bias in messages. We can see for example, it is to compare you with your grandmother, than the way that you express yourself, it's completely different that your grandmother typing messages to you or your family. So we have the same problems in region. So this is the most difficult problem that we have right now. Sometimes we try to, it's hard to say overcome, we try to overcome the situations. For example, I want to do some tests in Swedish. Is it simple? Because we do have people on swedish? But let me imagining Russia? I don't know if you do have employees in Russia? I don't know. But imagine that we don't we don't have any Russian employee. What we do, usually I try to translate Brazilian messages or try to translate English messages to Russian. But if you see that is not right. I cannot train algorithmic in your translated message and say, Oh, that right now I can detect russian, I know that I am lying, because we don't have real Russian messages to work on. So far as collecting data is the most difficult part. We can do some analysis after the in the modeling part, the machine learning part. But collecting data, it's really difficult. We don't we don't have some easy metric is or is a statistical things to analize that we can do a how analysis. For example, as I said, in country, I don't know how many resident messages I have, how many North America messages I have, or whatever. But it's still still how we are we in superficie. We are not going deeper. You know, we don't really understand how many I don't know youngers Oh, old peoples Do we have typing our message? So it's tricky. And I don't have the answer. I guess we we still do a how analysis. We need to go deeper. But it's really complicated to do.
60.	Speaker 2	Yeah. And will you say you are I know, you're in the other side. So you're in the machine learning side? So how will you say you actually in a very easy way, because we're not focusing that much on algorithmic bias? But like, how will you say you identify bias in your algorithm? So?
61.	Participant 6	Yeah. Okay. Let's put a concrete example. So let's go back to classification problem. So I have two class, we can quantify how how how's the tense of the average how it's going to one class or to another class, this can happen for multiple problems or for multiple reasons. Sorry, it could be because of the data. Or it could be because of the variety of the messages. For example, if you have more message for trying to buy things, then over customer service, of course, they're good, it's you tend to buy things because we have more data and indicating to this class. But imagine that I have the same amount of data for buying things and customer service. And I create a model to detect, we can always for example, in word level, I can see what are the words more importance for each of the classes. So in this case, I can indentify, for example, if my model is reading the whole message, or a specific parts of the message, in some time , another example. I read this one time ago, and it was really funny. They work at oh my god, animal classification, to detect dogs, cats, whatever. And they saw that to detect the Horse. The most common feature was the nose. So it's something that he's not detecting the horse. It's only looking then nose of the animals trying to identifying them. So you can see that is not using all the information of a label to really detect the things so we can check that in different ways. Again, yeah, it's we don't have One way to do that, it's you know, you are a

		from academically, usually we have 1000s and 1000s of papers and metric is and machine learning stuff is to do interpretability. But usually we do how, again, how analysis, because depends off the time that you have to work on this project. So usually we do, for example, if I have one month to develop a one time project, I you do, I will do some analysis. But you'll be, of course, on tools certain limit because I can balance the data. I can check the words, but of course, we can do more. But at least the how analysis we do, we have to do? Well, so my boss who won't approve the project **laughts**
62.	Speaker 2	is good to know. It's good that you do those analysis then. And then we have one last question. And then from there after we got like two or three specific cases. So this question we can take a quick so its about team diversity. Do you think having diverse teams have a positive effect on detecting bias? Do you have the diverse teams to in your?
63.	Participant 6	Yes, actually, I love that. to answer your question. Yes. Diversity is important. Detect bias. and yes, your team is really diverse. I live in Brazil, I have we have more two peoples in Brazil. We have three peoples in Sweden, we have a two people on India. And we have a pretty balanced the man and woman. Just let me check. I guess, oh, I cannot do the math on my head. I gotta not remember all the members. Sorry about that.
64.	speaker 1	That's okay. We just wanted an overview.
65.	Participant 6	But But of course, it's still really balanced. We have some oh my god, I forgot the word in English. It's sexual orientation, diversity as well. So we have gays and trans people. So it's, it's really interesting.
66.	Speaker 2	we feel that different kinds of people might be looking at that different problem in bias and the system, so that can help a lot. And so we're gonna continue with the I'm gonna share my screen. And we're gonna provide you with three specific cases. Can you see my screen? ** participant agrees** , perfect. So we have identified that bias could happen into a specific places. So it could be data bias or algorithmic bias. And that bias involves data collection and preparation, algorithm bias involves more like model development, model evaluation, post processing of deployment. And we understand that it might not be just one specific, one could be towards the other one, like maybe a lot of algorithmic bias happen because of the data bias. So we're gonna provide you with samples and you can let us know where a Where did the bias, where do you think the bias happened? And so, if you need to read what everything means, I imagine you know, but just in case like you just have it here, let us know and then we can use
67.	Participant 6	**participant reads** Actually, I found that interesting. I have never think about the model deployment. Can I print screen just remembered the names
68.	Speaker 2	We can send you the article that we use for this
69.	speaker 1	Yeah, I think we have them on the next slides as well. Yeah. So, first case, I don't know if you prefer to read it to yourself or if you want me to read it out loud.
70.	Participant 6	I can I can. I can read

71.	speaker 1	Please let me know when you're finished. Okay, thank
72.	Partici-pant 6	**participant reads**
73.	speaker 1	So then the question is starting from the data collection to the model de-ployment of the AI system, where would you say is the highest risk of bias occurring? Where did this bias happen?
74.	Partici-pant 6	I just say in the data collection
75.	speaker 1	Yeah. If we go to the next slide, we have some additional information, so I guess you you're still going with the data collection? Or did you?
76.	Partici-pant 6	Yeah.
77.	speaker 1	Then let's go to the next case. The question is the same for all cases. So you're gonna identify where the bias occurs. So you can read and we can talk?
78.	Partici-pant 6	Okay, I finish.
79.	speaker 1	Yeah. So what's your opinion? Where do you think the bias occurs?
80.	Partici-pant 6	I can, I can only say if it's data collection, data deployment and model de-ployment, all of them.?
81.	speaker 1	You can choose one, whichever you want.
82.	Speaker 2	And you can choose if you feel that there is a new one. .
83.	Partici-pant 6	Okay. Because it's, it's hard to see it, it has only one part. You know,
84.	speaker 1	you choose multiple if you want to do that as well.
85.	Partici-pant 6	Okay. Of course data collection, it could be a bias because we we have his-torical problem in this this case. Oh, it's really definitely case. Data Prepa-ration I don't know. But the Model development Yes. Use factor as age and criminal history. Doesn't feel fair for me. Model evaluation No. Post pro-cessing? What deployment is most difficult to save? Yes or no? But I would go with collection and I mean Model Development by the attributes use it to deploy a model I don't know if interpreting that but the attributes use it the model
86.	speaker 1	Okay, so if we go to the next slide, there are some additional facts so after reading this, then you can give us your opinion and your thoughts
87.	Partici-pant 6	In some cases, you have some specific information and another you have I don't know group information
88.	speaker 1	Yeah, it's like how detailed something
89.	Partici-pant 6	so they have data, the same data with different granularities Yeah, exactly. Oh, strange. But it happens sometimes
90.	Partici-pant 6	Well, it's more society's problem instead of data or Oh, keep the same data collection model development. It's the, I guess the most of the problems re-garding the data collection. I am bias you know *laught**

91.	speaker 1	I mean, we're not allowed to talk but we're happy with your answers. So, then we have one last case. So yeah, the question is the same just whenever you're done, let us know what you think. Okay.
92.	Participant 6	Okay, in this case, specifically in this case, I choose data preparation and model evaluation. Because we can, we can somehow check that in the model evaluation, but probably some data preparation can also be
93.	Speaker 2	I do have a question that are actually and so, will you say that you can identify, like, if there was data collection bias, you can actually identify that in the, in the steps like either mother evaluation or post processing, or the model. Or it's hard to identify,
94.	Participant 6	I guess it's hard to identify regarding the data, so that, but some examples you can use, financial lanes, if I am using age, or our gender to lend money for someone, I can clearly see that the problem is in the data. So I can see that the problems is is there. When I mean the data collection, I think more in the representativity of each group. So if I have, do have a lot of young people that that try to borrow money and will people that tried to borrow money or woman and man or I can see the these in the data collection with a simple statistical cell representative, the how much, man and woman I do have, but I'm imagining that if I have the same number of woman and men where's the problem of this? This part? Maybe I cannot see the data collection. But I can see in the data preparation so I can see. I don't know. Imagine it that I have. I have more men clinking in the ads. Opposite of the woman actually we do have a lot of, A lot of interviewing analysis that showed that men try to to apply to interview with less I guess 50% of the requirements and woman try to apply for an interview of over 8% of the requirements. So,
95.	Speaker 2	I read that too
96.	Participant 6	I imagined that problems in our culture. So, if we have more men clicking on these ads then the woman we can we can check that the data preparation somehow and I mean in the model evaluation, it could be the problem too because I can double check this data preparation in the model evaluation Okay,
97.	speaker 1	thank you so now this is like the second part, so reading this do you feel like you want to change your answer or you're happy with your previous answers
98.	Participant 6	**participant reads**
99.	Participant 6	**reading*** In this case, most of the companies you want to create ads for men? Data what is the definition of post processing? I forget.
100.	Speaker 2	So I don't have it in the top of my head right now. But it's basically when they see that the model evaluation is not going on the right way. So then you do like a little bit more of preparing the model from the model perspective, rather that other perspective So i think is more towards the classes
101.	Participant 6	**participant reads** so using this the first paragraph, I keep with the data preparation and model evaluation. But regarding the second part, I see Model Deployment
102.	speaker 1	so more of an algorithmic bias ? **participants agreess**

103.	Speaker 2	yeah, I I think there's they're very different in the sense like they don't do the articles that we read, they don't know either.
104.	speaker 1	like authors who investigated this presented this as possible reasons they couldn't determine which one it was either.
105.	Speaker 2	But yeah, of course, like these two, like these two points, they relate to each other in the way that the algorithm learns from patterns. And then of course, if men click then it cheaper to advertise for men, because those are the people that click but it's more expensive to advertise for for people that don't really click on the advertisement and which is female. But yes, this is this was it.
106.	speaker 1	Thank you for your time and for participating giving us some valuable insights. We really appreciate it.
107.	Speaker 2	Yeah, did you have any comment?
108.	Participant 6	I have a request for you guys. Please share with us the results after your thesis or invited us to to the presentation was
109.	speaker 1	Wow. We don't really know probably, I don't know, we haven't received any information about it yet.
110.	Speaker 2	We have face to face, but I don't know if it's, we can we're gonna publish the results and we're gonna publish the research but I don't know if we gonna do any online presentation.
111.	Participant 6	It's vary from university to university, someuniversity doing online and other face to face, actually face to face is better.
112.	speaker 1	It is.
113.	Speaker 2	I think so
114.	Participant 6	I usually it's at least me I felt less nervous. Because we have a small group, you know, you are, you're seeing a professor so you can see his or her faces, you know, I didn't get what you're saying.
115.	Speaker 2	Yes, but I will send you the the article that we use, because you asked me a few questions that I know this article has all the answers. So I will send that to you. And then you can have a look as well. Perfect, but we're super happy. Thank you so much. And then we will send you all the your transcripts. We're gonna put your name anonymous and where you work. I know you didn't say it. I think so. But anyway, if we find something that is that is regarding your persona, we will anonymize it.
116.	speaker 1	Okay. Thank you.
117.	Speaker 2	Thank you and have a nice day. And we'd like to know how it goes.
118.	speaker 1	Yes, thank you.

Appendix 14: Transcription - Participant 7

1.	Speaker 2	We are from Lund University, and we're currently working on a master's thesis, which is bias in the context of artificial intelligence systems. And we want to focus on a data perspective. With this, we want to check the effects of data has on the outcomes of the systems. And we want to identify where in the process ehh bias are more likely to occur. So where in the process from like you collect the data until you are deploying your model. And so with that, I let speaker 1 to talk about your rights.
2.	speaker 1	Yeah, just some general information, you're allowed to skip whatever question you want, or stop the interview. And the transcript of this interview will also be sent out to you so that you can go through it and make changes if needed.
3.	Participant 7	Cool
4.	Speaker 2	So with that being said, we can start with our first question. Could you tell us about your experience working with data and artificial intelligence?
5.	Participant 7	Sure. I have a background in both artificial intelligence and political science. So my interest has always been in the intersection of technology and specifically, AI and and policy or governance. I've done a master's degree on technology policy. And specifically, I've looked into operationalizing principles of ethical AI. So looking at how can you operationalize in an organization, ethical principles, and how can you make models, AI models adhere to these specific principles? And on this, I've done two extensive research projects. The first one was really more on the governance level. So looking at, within an a complex organization like a public ministry, how can you in the governance and policies of an organization make ethics work? How can you embed it in an organization? And the second one was more technical that was looking at if you have a model, for example, for hiring or for any other type of AI application? How can you set technical standards of bias and fairness metrics to certain performance requirements? And what would the impact be on the model lifecycle, on the implication, the imposition, sorry, of performance requirements on met of metrics of bias and fairness?
6.	speaker 1	Thank you. Interesting.
7.	Speaker 2	Yeah, thank you. And that is actually, I think, a great insight were in our research to get. And so I'm going to start with the topic of artificial intelligence. And we define artificial intelligence as the ability of a system to identify, interpret, make inference, or learn from data to achieve predeterminate organizational and societal goals. So this is a very general, and we will like to know if you agree with our definition, or would you like to change or remove something? And I can add the definition in the chat, if you want to read it again?
8.	Participant 7	Oh, yeah, that would be appreciated. In general, in general, I would say that, I think of it as descriptive and prescriptive systems of AI, where... where prescriptive systems really do kind of predictive work. So they

		look at training data, and they tell you on a certain new piece of data, what the best course of action would be, for example, does this client, will this client have a fraud risk? Should we take this client because of fraud or not? So like risk classification, and the other would be more descriptive. And that's more like on... on existing data, learning about patterns within your... within for example, for a bank if they have a lot of transaction data, or if you have a lot of hiring data, what can you see and learn from this group and this body of, of knowledge without making per se, predictions about any other future piece of information?
9.	Speaker 2	Okay, and thank you. And I think that is a great insight. Because yeah, the one is like finding new patterns. And then with the data, and one is actually just getting your patterns. And so with that, I would like to continue with the second question, which is like the literature that we have read points out that there is actually a need of addressing the dark side of artificial intelligence, which is here is the reason of AI systems, acting bias. So we are wondering, of course, we you work with ethics, but is this something that is talked about at your workplace or among your colleagues, ehmm, yeah?
10.	Participant 7	Yep, for sure. The most common way that we talk about it, is by looking at the entire model lifecycle, so from ideation, to Yeah, the building the model, then testing the model, training the model, putting the model into production, and then even the end, like retiring a model, all those different stages can introduce bias. And it's important to look at the entire model lifecycle to understand where model can, where sorry, bias can come in, because it can come in at any stage. And not just for example, what a lot of things have been written about is about the training data. So using historical training data has, for example, a certain bias, like the hiring one of Amazon is very known. In tech scene, there was a lot of white men that were hired. And of course, the model then preferred white men. So that is a very typical one. But bias can come in also in many other places, it can come in from the beginning of just thinking of an idea of a digital solution. So for example, if you think about just one example that I've that I've used is you have these days in supermarkets, a lot of these automated checkouts and just thinking about a digital solution, for example, automated checkouts can have a certain bias, because if you are from a background where employment is difficult, or it's hard to find jobs, you would not simply think about implementing a automated checkout system where people lose their jobs, you would think about, okay, what are we going to do with these people? Are we going to retrain them, you would have another idea about this digital solution than just thinking it's a great idea to make the business more efficient. And we're going to implement this digital technology, because it's innovation and it's great. You will have a another perspective on it. And that's already where you see bias towards using a technology or not.
11.	speaker 1	Great explanation. Thank you. So that
12.	Speaker 2	yeah,
13.	speaker 1	yes

14.	Speaker 2	you can continue.
15.	speaker 1	Yeah, that brings us to our next section, which is more focused on the data. So my first question is, when collecting data, do you have any specific requirements that you follow or any other things you consider?
16.	Participant 7	Well, so I don't make models myself. But if I would be talking to a client that is making these kinds of models and is collecting data, and I would be looking at that, in terms of data collection? I would, um, I would always um, let me think about this. I mean, I would always say, first of all that, especially for data collection, it's more of a privacy question on top of my mind. So GDPR comes into play. What is the purpose and scope of this data collection? Is there a principle of data minimization? Are you just collecting for collecting sake? Or do you have a specific purpose for what you're collecting it? And are you using the least amount of data that you need for the purpose and the goal that you have in mind? So that's really important. I think maybe another aspect which is related to that, that's data labeling, that's where you see a lot of bias creep in, when training data for an algorithm is created, that the algorithm is trained on the model is trained on that people, usually those are like AI, people that studied AI, there's a lot of them these days with tech companies that are just used for labeling data, and their only job the whole day is labeling data like this is a cat, this is a dog this is this this is that to label data. And there you see that you can quickly get a kind of bias, if you have only have white men labeling data, you are going to get a bias in the outcomes. So then you would be the question would be, what does the team look like in terms of diversity? That is doing the labeling? And do you have somebody from various backgrounds and various cultural, ethical, etc? Even academic backgrounds that are looking at the data? And is the interpretation similar amongst all these people? So if you have people from I don't know, I'm just saying something random 10 different kinds of ethnicities and cultural backgrounds that all agree this is a cat, then you're much more likely to be correct that indeed, it is a cat than if only one person is saying, yeah, okay, this is this is a cat. Well, for somebody else, maybe it's something, another animal. It's a trivial example. But what I'm trying to say is that the diversity of the, the, the increasing the diversity of the team, prevents you from having blind spots. And thereby discrimination in the end.
17.	speaker 1	Thank you, so a follow up question, do you work in a more precautious way when you're dealing with demographic information? Or are the requirements same? And you work pretty much the same way as you do with all types of data?
18.	Participant 7	No, I mean, that's more of a legal question. I would say, I think because it's, it's a GDPR thing. There are protected classes that are defined within the GDPR. And those protected classes require more, have more rules on them on how to treat the that information. Um, I don't have a law background. So I don't know those exactly by heart. But it is things like religion, date of birth ehmm I don't know, some some some of these. Yeah, sexual orientation, etc, those those kinds of data points are protected. Yeah.

19.	speaker 1	Thank you.
20.	Speaker 2	And I can continue to the bias part. So we're wondering, like, how do you work in order to detect bias in your systems?
21.	Participant 7	***Laughing*** So you can do that, in two ways. I would say, the first one is a more technical one. So in that way, you have just technical metrics of bias and fairness. So there, you have these umm, there's this famous article, academic article, which defines 21 Different kinds of fairness, like mathematical definitions of fairness. And you have technical tools, this, these days that are developed, like one is called what's it called, again, Microsoft made made it it's called fair something, I would have to look it up in a second. But it's like a technical tool, you can put the algorithm in it, or the sorry, the data in it. And it can tell you if there is bias towards certain kinds of classes that you've defined in your data set. So you use this tool, and the tool tells you, okay, it's kind of Explainable AI. That's basically what people say about Explainable AI, that you understand what the data looks like if the data is balanced, if there is not certain classes that are more predominant than other classes. And if it's, if it's, if it has bias or not, and you can fix it quite easily that way by by making it insightful and understanding the data. Another way is more I would say, more outcome based. So when you use a model, where you use it is very, very important. And I think this is something that the let's say the Explainable AI community is often missing, because you can understand a model technically really well, you can, technically looking at it from the inside out, you can see like, okay, it is very balanced it is it has no bias, it is completely fair. But then if you use it or fair in the definition that we want it to be, then if you use it in a context, which is very unpredictable and changing and dynamic, then it can have completely different outcomes. So, for example, one thing that I was looking at right now in the financial sector is there is a bank that is using algorithms to detect money laundering in Eastern Europe. Umm, the model, the model is very explainable the data they used is they're using, it's very balanced, it has no extreme biases. But using this model in Eastern Europe now means that you are inadvertently targeting war refugees. So that means that the way you're you're using it has bias in the sense that you you have unexpected discrimination towards people that you do not want to target. Because there is no way that in your model, being a war refugee is defined as a class. So then you can understand the model within almost all circumstances. But then there's this one other extreme circumstance, because there's a war in a certain region, which means that you're suddenly targeting for your anti money laundering model, you're targeting war refugees and making them their life maybe even more difficult because they're unable to open bank accounts or other reasons, because because they're flagged by the system. So that's what is in the academic literature, often called the socio technical environment. And that you need to understand not only from, I would say, the Inside out of the model, how it works, but also from the outside in how it is applied, and how it is perceived and used by people in the actual world.
22.	Speaker 1	Thank you. Great insights.

23.	Speaker 2	Yeah, thank you so much. And I, we love examples. And that was great. And we have one more question here. And then we go towards the cases. So based on your previous experience, what is the most common type of bias? Or where can we find bias the most?
24.	Partici-pant 7	That's a difficult question, the most common type of bias, umm.. Yeah...
25.	Speaker 2	You can mention a few, doesn't have to be one.
26.	Partici-pant 7	No, but it's it's like, I would say um, but, but it's, it's maybe not exactly what you're looking for. But it is something that I see very often is that there is a bias towards the umm the unassumed, or the assumed the the the kind of the always assumed benefit of using digital technologies. And very, very rarely is the question asked, Do we, we have a problem X is a digital solution, the best solution to solve problem X. People always think about, OK, we are using digital solutions. We're going to solve everything with digital solutions. And it's much more important. And there there's a strong bias, I think, in the assumption that digital solutions are always the better solutions. I think it's it's a lack of creativity, that for whatever problem, digital solutions are the best solution. And I think it's important to take a step back and to look at what is really the problem? And do we actually need an AI solution for this? Do we actually need a digital solution in the first place? Or are there other ways that are more human centric, in that sense, than than a AI algorithm?
27.	Speaker 2	Thank you. And that's a, that's a great approach. And that's a good insight that we're going to put in our research. So X, could you share your screen and then we go really quick to the cases.
28.	Partici-pant 7	Yeah.
29.	Speaker 2	Yes, so here, and we have identify, like, these are, this could be like the specific place where bias could happen. And so we have identify like the process which is could be like data collection data preparation or in the algorithmic bias will be like model deployment, model evaluation, post processing or deployment. And so with this, we will go towards the cases. And then in the cases you can identify it. We know that it's like a lot of factors around. Maybe you can choose one of them or a few of them. And if you need to know what each one means, then we have like a specific text down ehmm, that you can read. Otherwise, we continue.
30.	Partici-pant 7	Yeah, please!
31.	Speaker 2	Thank you.
32.	speaker 1	So this is the first case. I don't know if you prefer to read it to yourself, or if you want me to read it out loud,
33.	Partici-pant 7	No, I can read it, give me a sec.
34.	speaker 1	Yeah!
35.	Partici-pant 7	Okay, yeah!

36.	speaker 1	Yeah. So starting from the data collection, to the model deployment, where would you say that the bias occurred for this system?
37.	Partici-pant 7	Umm, well, it can be in different ones um, let's see. I would say it's either data collection or um the post processing or evaluation.
38.	speaker 1	Okay, thank you. So if we go to the next slide, we have an additional fact.
39.	Partici-pant 7	Yeah!
40.	speaker 1	Reading this, you want to change your answer?
41.	Partici-pant 7	Yeah, that it's definitely data collection!
42.	speaker 1	Yes. Then let's go to the next case. The question is the same for this one.
43.	Partici-pant 7	Yep!
44.	speaker 1	So just let us know when you're ready.
45.	Partici-pant 7	So um.... this is actually, I know this this case reasonably well. And um this was, I think, this had to do with the ehh damn, I don't remember exactly, but I think it was the the model development, because in this case, the uh it was the problem here in this case was the fairness metric that was used, because the interesting thing is that, as I said, there are 21 different different kinds of ways to define mathematically to define fairness. And many of them or not, most of them are mutually exclusive. So in this case, one fairness metric was used. And the people that analyzed this case said that it was biased in another fairness metric. But it is impossible to have it balanced in both fairness metrics. So they had to make a choice. And that choice was made in the model development that this fairness metric they are going to use. And so it.. result, the result was on one type of fairness metric that it was dark skinned people were more often classified. But on another it was equal to white people. And, and that was the difficulty in this in this case.
46.	speaker 1	Thank you. So going to the next case, we have some additional information, but since you were invested in the case and have knowledge about it, maybe we can skip that due to time and just go to the last case.
47.	Partici-pant 7	Yeah
48.	speaker 1	So this is the last case.
49.	Partici-pant 7	Oh, interesting. Okay. Uh, yeah, I would say this is um, I think this is... Well, it depends. But let's say based on purely what I'm reading here, I would say model deployment.
50.	speaker 1	Yeah. Okay, so then the people that wrote about this case, they don't really know, either, but here are two possible reasons for it. So we would just like to hear your opinion on them. ***Showing the additional fact***.
51.	Partici-pant 7	Yeah, yeah. I mean, it's it's um... I would say this is a kind of like unsupervised learning um... So it's just going at it and seeing and wanting to

		maximize um... clicks, and so on certain content, men have higher, higher clicks. So then, obviously, the model focuses on men. One interesting example about this is that I don't know if you've heard about, or actually, no it was not Cambridge analytical, but the Trump campaign when the the the social media wizards that did the Trump campaign, he did an interview where he explained that they used AI models, and they had used of of the same ad, they had used 1000s Different versions with like, different shading of color, different wording. They had changed everything to the smallest details. And they thereby maximized how certain um... how certain groups how certain demographics, would click on the ad. So for example, if you had a buy now or donate no button, for Asian Americans, they would be more likely to click on a green button than a red button and women would be more likely to click on I don't know, whatever, a pink button, I'm just saying something
52.	speaker 1	Yeah
53.	Participant 7	but so... so in that sense, it really depends what the ad looks like. And if you are using one, one type of ad, and women are less likely to click it. Is it the ad? Is it the content? You cannot answer that question without trying out different versions and seeing how you how you tailor it to to certain demographics. Um, let's see the second one. That was sorry, I was still reading it. *** reading*** Yeah, I mean, that that is also part of, I guess what I said? It's, it's, it's, it's making the most cost efficient choice. It's making not so much the most cost efficient choice. That's, I think, a different wrong, use of words, it's making the highest... umm, what's the word I'm looking for? The usage and marketing...
54.	Speaker 2	Profitable?
55.	Participant 7	No, not profitable, but the highest click rate basically. So ehh the highest follow through. That's what it's optimizing always for it's optimizing engagement. And cost efficient, is it might be more cost efficient, or more or more profitable. But as a more complicated question that has more factors attached to it. The model in the essence is just there to maximize engagement. It's maximizing clicks, comments. Um... Yeah.
56.	speaker 1	Thank you.
57.	Speaker 2	Yeah. We are super happy about this interview. We wish we have more time. And we know you're a busy man. So we won't take much of your time and we will send you this transcript and we will anonymize all the personal information. And if you have any comments, or any insights that you would like to

Appendix 15: Transcription - Participant 8

1)	Speaker 2	So I will give you a quick introduction of our research. So we are two students here in Sweden in Lund University. And our thesis and the name of our thesis is bias in the context of artificial intelligence systems. And we want to focus on that perspective. And so tackling these from a data point of view, so we're interested to actually find how that affects the outcomes of the systems and identifying where in the process the bias can occur. But of course, we know that there is many factors around this. But we can always like hope to find this.
2)	Participant 8	Yeah, by the way, it's a very cool field to be in.
3)	speaker 1	Well we are very excited about it too. So we are a bit tired now like looking forward to being done with it. It's been exciting to learn about it. So yeah, I'm just gonna give you some general information. The interview will be about one hour, and you have the right to stop whenever you want, or skip whatever questions you want. And as we said, we're recording and then we will transcribe it and we will send it out to you. So you will have an opportunity to go through it and make changes or add or delete whatever you want. And so yeah, with that being said, we can start with our first question. Could you please provide us with a quick introduction about yourself?
4)	Participant 8	Yeah, for sure. So my name is X. I grew up in Y, I lived in Z for the last 10 years. My background is aerospace engineering. So my background is not machine learning. You know, it's a new field anyway, academically. But I've always been in love with mathematics and logic and programming. So naturally, companies, when companies started adopting more advanced form of automation, like AI, young people like me, in technical roles were pushed towards those teams. So I found myself in data science to three years ago, it's a natural progression. And now my focus has been NLP natural language processing texts. My job before this was working for a actually big energy company in Z. So the focus was more corporate applications for engineering. As of February, I moved to Q, as head of the machine learning group in company X, which is a tech startup. And as you know, probably our focus is detecting bias in tax and offering inclusive alternatives. That's my entry. Great,
5)	speaker 1	thank you. Could you tell us a bit more about your experience working with data and artificial intelligence?
6)	Participant 8	Yeah, so I touched upon that a little bit. My first experience in data science was very specialized, quantitative engineering data. So highly not interpretable. And very complex. But my first experience with machine learning in this more, I guess, public form of it was building models for a big energy company. So we work with the marketing team, the sales team, build very customized models. What I'm doing right now is we're building a machine learning model, like I said, to detect bias and offer alternatives that creates an inclusive language.
7)	speaker 1	real interesting, thank you. Yeah,

8)	Speaker 2	thank you. And we really like that. Yeah, that you say like, it's, it's a new field. And it was a natural kind of change to actually go from engineering towards that, that scientists and thing as well like machine learning. And so with that, we want to start with like, asking you about the definition of artificial intelligence. So we have one, and we were wondering if you agree with our definition, or if you're going to change something or remove something. So to give you the definition, and AI is the ability of a system to identify, interpret, make inference, or learn from data to achieve determinate organizational and societal goals. So do you agree to this, or would you like to change something
9)	Participant 8	I agree, but it depends who you talk to. If you talk to a technical person like me, the last part to achieve predetermine goals, unfortunately, is not the focus of a technical data scientists. If you ask someone who's less technically focused, then they will add that as a as an important part of it. That's the end. Yeah. But I would agree. It's a good definition
10)	Participant 8	And you as a technical person. Yeah. What do you think?
11)	Participant 8	I am a technical person. But I'm trying to connect the societal part to this, which is why I joined this company, my previous roles, which is very data driven, in an abstract form, I didn't really, I couldn't sense what the goal was, here, the goal, concrete, we're trying to improve inclusion of certain minority groups in Scandinavia. So I agree with the definition.
12)	Speaker 2	Thank you. And that is true, like different kinds of people have different kind of definitions. And we really like we're social science, too. So of course, we're gonna like more of like a social approach, rather than a technical approach. And but to continue with this. And the leader to that will actually read, he points out that there is like a need of addressing the dark side of artificial intelligence, which is like the reason systems are team bias. So we know that your company, this is something that is talked about, but we were wondering, like, also reflecting from your past experience. And then this one's like, when was this something talked about before as well in your other companies? As well, like now, compared to the one that you're that you are?
13)	Participant 8	Sure, yeah, that's a good question. So I can tell you in my current company, because our product is all about detecting bias, it would be hypocrites if we didn't care about what's going on in our black boxes. So absolutely. We we talk about it. And we, it's a it's an integral part of our process, and machine learning models. In my previous company, where it was the focus was more engineering type of data analysis with big data that was complex in an academic sense. The models weren't always interpretable. So that wasn't the focus. And at some points, it could get out of hand, and it becomes a black box and dark by default. So yeah, there's a contrast between my previous experience and this one.
14)	speaker 1	Yeah, we read about that in the literature as well, like some corporations, they see the profit, and that's what they're going for. How do they get there doesn't really matter?
15)	Participant 8	Yeah, yeah, I can tell you that if we strictly rely on those lockboxes. Without us trying to understand them, we could generate 75% of the business value, probably. Wow. Yeah, that's really interesting. But that last? Yeah.

		So we could generate 75% without really involving humans. In our product, it's actually true. Can you can do the heavy lifting, strictly with mathematics, without me trying to look at the data and understand what it means that don't do that.
16)	speaker 1	And scary at the same time?
17)	Participant 8	Yeah. Yeah. It's one of your next question. So I'll hold off on that.
18)	speaker 1	Yeah. So going over to the data view, we have some data related questions. When collecting data. Do you have specific requirements? Or how do you like work when you collect your data?
19)	Participant 8	So because our products goal is to detect bias, we actually want bias data? Yeah, so our use case is different from other companies you talk to, we don't want to curate an unbiased data set, we actually want to study data. So we don't really make sure that we get an artificial representation. We want data that represents reality, whether it's racist, sexist, ageist, and all.
20)	speaker 1	Yeah. And are those open data sources or where do you get it from?
21)	Participant 8	So our data is currently or products analyzes job ads? Yeah, and those are public jobs. have us available on websites like the hub? LinkedIn?
22)	speaker 1	Yeah. Thank you. So that brings me to my next question, which is how do you ensure that the data you use is of high quality, which may be in your case doesn't really matter?
23)	Participant 8	Yeah, like I said, we want the data as it is. We want to study how biased it is. The quality parts, I'm not sure how you define quality, if you mean from a machine perspective, versus a sole social perspective.
24)	speaker 1	Yeah, like when we talk about high quality and like, based on the literature we're read, we're mainly talking about, like making sure that it's representative like doing pre processing and stuff like that to make sure that's okay. Suitable.
25)	Participant 8	So in that sense, now we take it as it as it is, because we're interested in the the low quality part of the data. Yeah.
26)	speaker 1	Thank you. And then, like, some of our questions are not that relevant, so I'm going to skip them. But would you say that one can ensure that there is no risk of bias in like regular systems? Based on your experience? I know you touched upon it for the black boxes in your previous company.
27)	Participant 8	So to answer that question, I have to break it down into three parts. There's the data, which is the inputs, the models, which I think you call algorithms in your language, and then the orchestration of everything, which is the system. So do you want me to answer based on our company or overall,
28)	speaker 1	overall, like your experience? Yeah.
29)	Participant 8	So if we keep that breakdown in mind, so with the data, you know, that you have to make sure that you get representative data, and preferably balanced, your data should reflect the group that you're studying? Right? Like, you know, about the famous Amazon use case where? Yeah, it was biased against women. Yeah, that was because they didn't have enough sample points for women. So the model didn't understand that a woman could also be good professionally. Yeah. So the high quality representative balanced data, that's the first step. And you feed that into the models. The models

		that we use in data science can be incredibly simple, just statistics, or they could be something a lot more complex, like, have you heard of Bert? language model?
30)	speaker 1	No, no.
31)	Participant 8	So there are some very, very powerful language models. These are universities that worked for years to create these language models that mimic the understanding of a human to certain language. And those actually very few data scientists understand, though, that's a black box to someone like me. And the danger of those is, they're open source, so anyone can use them. But we don't know where they got their data from. But they are incredibly well established. So everyone uses Bert, you're gonna hear about BRT. But if you ask, have you actually looked into it? Chances are everyone's gonna say no, because it's too complex. Yeah. But the result? And
32)	Speaker 2	that's scary, I think
33)	speaker 1	it's real dangerous.
34)	Participant 8	Yeah. So models like Google Translate, use fancy language models like that. They're very powerful. And they work. It's just we don't know what's going on inside. And they keep iterating on them to improve them. Like you notice Google Translate 10 years ago was horrible. And look at it now. Right? Yeah.
35)	Speaker 2	I remember like, we were going to touch up on this like last year, and I was saying, like, I speak Spanish, and my language is female or masculine, and I was in Google doesn't have it. And then when I went to check, it's like, yeah. Oh,
36)	Participant 8	they have? Yeah, it's I think they added that recently, I just thought, yeah. So talked about the data, the models, and then the system is how you add an element of human validation to interpret what's going on. And it's nice to have that when you collect the data, when you feed it, the black box itself, so a data scientist to understand the mathematics. And then the last step, which is actually quality checking the output of your model that's on the system level, you need human validation and checks. And there's also a technical term not sure you heard about it, it's called adversarial testing. Have you heard of it? No, not really. It's very cool. But think of it like, like an exam that's designed designed to fail the student. So it's like a set of questions that we ask the model to try to break it. And in the example of the Amazon recruitments model, the testing would be if we created a data set, which just women very competent, it would absolutely fail. And we would have spotted, who would have spotted the output, if someone had tested the black box with just that data. We wouldn't have had those problems.
37)	speaker 1	This is amazing. I'm learning so much stuff right now.
38)	Participant 8	I have to know these things.
39)	speaker 1	That's good.

40)	Speaker 2	Yeah. And I think that's great that way that you actually also put a good thing into example. And, and because it's something that it helps us, contrast with others, and then trying to understand, okay, this is the theory, but then the practice might not be a similar as the theory.
41)	speaker 1	Yes. So then our last question for this section. It's not really relevant to your current job. But I don't know if you work with demographic data previously, if you did, did you like have special routines for how to deal with that? Or you just manage it like all the other data points?
42)	Participant 8	I worked on that very briefly with some human resource data. And the complicated part is protecting the data. We have to anonymize it. Yeah. Which is not as simple as you think. Where it's not a matter of just hiding names and numbers. Sometimes you can trace the person with certain hidden attributes. Yeah. And those are very hard to simply just hide, so anonymizing and that's based on the regulations of each country, so Canada is a lot more regulated for data protection than then Europe. Oh, really? That they're a lot more strict. Yeah. Yeah. For the rest, I don't really have good insights to share. Sorry.
43)	speaker 1	No, this was really good. Thank you. Um, then we can continue to the bias part, maybe? Yes. Yeah. It's
44)	Speaker 2	that point that you used to touch upon? Yeah, we read. And then other people has shared that sometimes, it's not enough just hiding an attribute or hiding a value because it is embedded in something else. For example, like there is like correlations around so like, doesn't matter. If you delete one specific, like skin color, there is still like in the address where you collected the data, there is this distinct, distinct color. And, but to continue with that, like the bias point of view? So we are wondering, like, of course, you're working in a system that detects bias. So but maybe to continue with that question, though, how do you make sure that even the system that you're using, that you are creating to detect bias, it is not bias?
45)	Participant 8	Yeah. So when you sent me the guide that I read into it, I was a little confused, because there's like a nested concept of bias. There's bias and in our in AI, and there's bias in texts. So remember, how I broke down the workflow, data model system. So in my team, the way we ensure that there is no bias is at the model and system level. So at the model, we make sure that the, the mathematical libraries that we use are simple and interpretable. And if I open the mathematical equation, and I read them, I can understand them. So all our models are like that, except for Bert, which I mentioned. And that's very powerful. And like I said, we can't Understand that black box, but the way we mitigate that risk, and that uncertainty is we take the output of that model Bert, we take the results, and we QA them by a team of experts in linguistics. So we don't, our default assumption is that Bert will be bias. We have a team of five linguists whose job is every month on a regular basis, we give them a list of 1000s of sentences on Excel, and they'll go through it and confirm this, okay, it's not okay. And it's very expensive for companies, especially like us. Like I said, we can get to business value 75% of it without involving these people. But we choose to pay five people full time, that have master's in linguistics, to look at that output, and make sure that that black box isn't racist, sexist ages that have the DATA step. We want the bias that we're not really processing that the

		model part I talked about it. And the system level is how we integrate that manual review into our process. And like I said, expensive and timely,
46)	speaker 1	Yeah, but we're very thankful that you're doing it.
47)	Partici-pant 8	I mean, the mathematician in Me thinks 75% is good enough.
48)	Speaker 2	Yeah,
49)	Partici-pant 8	human in me is that I'm a data scientist, but I'm skeptical of automation. So
50)	speaker 1	yeah.
51)	Speaker 2	I think we, we need that. Like, it's not just, you know, it's not just technical, we need to like, point ethics. Because otherwise, like us, like the three of us, my might be affected by this bias. And we were actually checking like, one, one interview, we checked, like, are you scared of this and the interview with a white male? He was like, No, I'm fine. I think there's a profit. So it's like a contrast, because of course, if it's not affecting you, then you don't realize that it's actually bias happening. And but yet to I guess, to continue with this, let me see how to, because all the questions were for were general, so to allow us for cooperation. So what kind of biases have what do you think is the most common type of bias? Like which bias Have you identify?
52)	Partici-pant 8	You mean, in our company or my experience in life?
53)	Speaker 2	In both?
54)	Partici-pant 8	Well, obviously, as a woman in science, I can tell you sexism, and ageism are a big deal. Because I'm a female and I'm young,
55)	Speaker 2	we are work related maybe.
56)	Partici-pant 8	Yeah, for our product. So if you look at our software, we break down the affected groups in society into five categories. There is gender, ethnical, group. Age, gender, ethnical, group age, physical ability, and neural ability. So there are five groups of people. And our research shows that there is a lot more attention and emphasis on sexism and racism. But it also shows there's a growing, growing attention on neurodiversity, because mental health is becoming more important in people's life. So things that are considered mundane like ADHD spectrum. dyslexia are actually getting more important, at least in Scandinavia. Yeah.
57)	speaker 1	Well, that that is important because based on like, where we're taught, like, which country these things are happening in like, they will focus on different things like some countries will not even acknowledge like ADHD is something like so.
58)	Speaker 2	And to continue with that, maybe only a more technical approach rather than a social approach, like we were looking more like causes of bias in the sense like historical, maybe historical data represents a sampling representation or things happening on the model itself. So which kind of bias the curve you encounter. In that sense

59)	Participant 8	in our model um, I guess the bias here is the lack of represent representation of certain of these groups like the neuro diversity, physical ability. So what we do, we scrape internet to get the job ads. And a lot of those job has used sexist and racist and ageist language. But those two groups are left out, because first of all, we don't really understand what offensive language is to those groups. So our data is not very well representative of those two groups, physical disability. So our models are imbalanced. In a sense, they focus more on sexism and racism, a bit of ageism. So we do have a bias. The way we mitigate that is, the five linguists that I mentioned, we have, they form a research team, half of their job is reviewing our data. And the other half is reading academic research on bias. And we get some insights from them. And we try to feed it into our model somehow. But we're still not fully there yet. We focus more on sexism and racism.
60)	Speaker 2	Yeah, thank you. And that is, I think, amazing insights that we can add. And so we have we have one more question. What is very like, I mean, the name of your company is. So, of course, diversity
61)	Participant 8	I can pull these stats for you, if you want.
62)	Speaker 2	No, I think it's like in general, like, of course, do you think having diverse things can have a positive effect detecting bias?
63)	speaker 1	I mean, we already talked about the white guy not being worried.
64)	Participant 8	*** laughing*** I'm surprised. I mean, not generalizing, but I do. My entire career, I've, I've almost been the only woman in my team, and the youngest. So I had to work really hard to prove that I was the technical expert, when like my results showed it. What I've noticed working in a team like this one, where actually our gender balance is flipped. We only have we have a few men in our office, it's mostly women. So it's a very, it's an interesting social experiment for me, having been the only woman before and now we're almost entirely women. But I can the perspective is broadened by bringing different people. And I don't mean just in soft sciences. Even in hard sciences. Yeah.
65)	Speaker 2	So but that's great. We got to the fun part now, which is I, I have just so many questions for you in general, but I think like we need to start these types.
66)	Participant 8	I'm just gonna say, if you want I can do a follow up if you want.
67)	Speaker 2	Okay, now, that's amazing. So we will right now, we will give you three cases. And I'm gonna share my screen. Let me just see. And then sure. Can you see my screen? Yes. So we have develop this model from the literature. We understand that in my head, like do some people might add something to people my and delete on specific parts. But what we have identified in the literature is that bias can happens from the data collection to the model the deployment. We had the bad days between like data bias, algorithmic bias, where we understand like they are interconnected, and one might affect the other one and vice versa. And so we will, we're gonna show you three different cases. And then we're wondering like, in the cases do you can identify by bias from your point of view, and then maybe later, you can let us know if you have any any comments in this model? If you

		feel like it's a good representation, one representative of what real life is, or if you need something or a script, okay, get so to continue. Yeah.
68)	speaker 1	Yeah. So this is our first case, I don't know if you prefer to read it to yourself or if you want me to read it out loud.
69)	Participant 8	Um, let me read it quickly.
70)	speaker 1	Yeah. No worries.
71)	Participant 8	Yeah, okay.
72)	speaker 1	Yeah. So our question is starting from data collection, to model deployment? Where would you say the risk of bias is like the bias that we see in this case? Where is it from?
73)	Participant 8	Definitely, the data that you use that they use to train the model didn't have enough samples of Asian ethnic groups with the specific AI features. So it's definitely the data collection, I would say, not representative. Yeah. And then maybe the model evaluation because the training sets didn't contain enough that I guess, also their tests that how they evaluated the accuracy of their model probably also didn't have enough diversity like this, that a problem.
74)	speaker 1	Thank you. If we go to the next slide, we had some additional information, but it just confirms what you just said. So it doesn't really matter. But their model evaluation point that you brought up was quite interesting. They didn't mention it in the article. So that's something that we can talk about. So then we can skip right away to the second case.
75)	Speaker 2	Yeah, I just have a question there, like, so would you say if, if data, let's say data is bias, but then that will be shown in the model part?
76)	Participant 8	Say that, again, if the data is biased.
77)	Speaker 2	So if the data is biased, this will be shown in the algorithmic part in the model evaluation for and then people will be aware that the data is bias.
78)	Participant 8	So if, in this example, let's say all the data they use were white people, the model, the algorithm actually, inherently is unbiased when you start, it's what you use to train it, the teacher that it's like an innocent child, that's our model. If you raise your child telling him that, black people are bad, they will think they're bad. But if you don't tell that to your child, and you just let your child meet different people and appreciate them, then the output of that child or model would be these people are not bad. So in this case, it's the training data. But my point is the if they had tested their model, with a different type of data, like adversarial testing that I talked about, they would have realized that something was wrong.
79)	Speaker 2	Okay. Thank you. And continue.
80)	speaker 1	Yeah. Okay. So this is the second case, the question is the same. Okay. So when you're done reading, we can talk I know,
81)	Participant 8	I know about compass. Yeah. Yeah, it was biased against black people, right?

82)	speaker 1	Yeah, exactly. Yeah. Yeah.
83)	Participant 8	I've read so much about this. And there was a man that sued them. And he won. Right? This? Yeah,
84)	speaker 1	I think so. It was a trial. And it was, it was a really big case.
85)	Participant 8	He kept applying for appeal many times, and he had an excellent record in jail. But it would never go through the algorithm. Yeah, despite him having all the required features except the skin color. So in this case, I know it's also a data problem. But then I, I think in this case, it's the model development because I followed that trial. And they brought in data scientists to explain the flow and they couldn't explain it. So the problem was data. I think, again, lack of representation of darks can keep With good behavior, and also the development of the model itself, it was proprietary, and they never could explain it. So.
86)	speaker 1	So would you say that that would be like a reflection of the developers bias?
87)	Participant 8	It's interesting. I think it's there's definitely the data problem, like in the first use case. But in the second one, I think it's the the model evaluation part here. Yeah. This one's what does the literature say?
88)	speaker 1	The literature has, we can skip to the next slide, maybe because we have some additional points. I personally think that the third one is quite interesting. But you can read them, and then we're interested in knowing your opinion.
89)	Participant 8	Yeah, that's what I was trying to say, by lack of data, representing black people with good behavior, you know, because statistically, they collect data from disadvantaged neighborhoods, that, that have a history of bad behavior, because no one wants to be poor, but poverty forces you to do certain things. So it just reinforces the things like the wealth gap, societal gap status gap. It's unfortunate. But the most interesting
90)	Speaker 2	and one thing that was just that was brought up by another an interview with that actually was fairness metrics problem, where they had two different cases where the model was performing really good for dark skinned people, but was not performing that good for white skinned people. So they had to, and then the other case was like the opposite. So they had to make they have to make a choice in the fairness metrics.
91)	Participant 8	Yeah. Yeah, I can I can visualize that. Yeah. But that's unethical, though. Whoever developed this model did it unethically? Yeah, because you can't choose when you're facing a decision like that. It's not up to you as a mathematician to choose what's best for society. You just do
92)	Speaker 2	it's sadly and we just say that even their skin color is removed. Remove that is still there is like that embedded into order and at attributes.
93)	Participant 8	I can't I have to look at how the features are represented. But if third point is specifically about skin color, then maybe removing it would unbias some of the model, then it depends on the data, so I can't answer.
94)	Speaker 2	That's okay. And we can continue with the last one. Yeah. Okay.
95)	speaker 1 38:48	So the same question here. What would you say is the reason for this?

96)	Participant 8	So, the question here is, how do they value who's interested in an in a STEM add or not? That's interesting. If they base this strictly on language, then that could explain why. So, ads are usually based on cookies, right? Is that part of the Okay? So, if they're based on the cookies, then they're probably
97)	Participant 8	they see an assumption that maybe women browse certain types of websites a lot more than men that fall in different categories like, like household shopping Again, it's an assumption. Yeah,
98)	speaker 1	we have some additional facts like the people who wrote this article about this case, they don't really know the reason, but they gave suggestions. So if we skip to the next slide, they gave these two points as possible reasons for this. So
99)	Speaker 2	And just to add for when you say, like, adds, in social media, they are often depending on the click rate. So as it says, In addition, and finance,
100)	Participant 8	this is actually exactly the problem that we are trying to solve. So our research here shows that women are less interested in STEM jobs, because the language is not attractive to them.
101)	speaker 1	Oh, I actually read like an article about that, like I think it was a month ago or something that just exchanging some words, in the job ad were like, attract a lot of women just by some words.
102)	Participant 8	And there's research that we have that proves that it's exactly what we're trying to solve with the sexism bucket. So I could if if this is about click rates, which I'm not expert in social media, but showing a job ad steam about that contains language that pushes the way women would lower their click rate. So we would explain this. Yeah. If that's the hypothesis here. What's the last one presenting us? Oh, I didn't know that. That was that. Is that a fact?
103)	speaker 1	Yeah, we didn't know that either. But apparently, like I think it was the age of 25 to 40 or something, if I'm not mistaking. Those are like the most expensive target groups.
104)	Participant 8	Do we know why that's fascinating.
105)	speaker 1	No, they have discussions regarding like, those are easy to pinpoint. And they like that age, people like to buy stuff, they really they don't really need and stuff like that. So therefore, it's like, something good for companies who want to make profit.
106)	Speaker 2	Also, I come from a marketing background. And I had done i hackery, that myself is a specific specific ads like this. And let's just say like, I create this specific different hearts, depending on who I want to target. And, of course, if they ad will be cost efficient, because if only mail is clicking, then it will be more expensive to target to somebody to target to somebody that is not clicking, and to show it to somebody that is not clicking. So there is like that's like a different way of seeing cost efficient. And as opposite as well like to have and like it is for sure that these specific groups are more expensive.
107)	Participant 8	So the question here is, why do women click less? Effectively? Right? Yeah. My theory working in this company is it's not a theory anymore. We validated that stem, Java language language is very actually repulsive to women, when you read it, and there's language that we want an intelligent

		programmer that can work in a fast paced, competitive environment, blah, blah, blah. To most women, it's too much and scary and intimidating. It's too masculine. The term is a agentic. Versus and it's a it's we validated that. So that would be my biased answer here.
108)	Speaker 2	But then come when you're comparing if you had to choose one here,
109)	Participant 8	between these two
110)	Speaker 2	Yeah, between that bias and algorithmic bias on being foreign somewhere
111)	Participant 8	not understanding the reason why these two are. I don't understand why these are true facts. But if I assume that they are I lean towards the first one. Because I assume that like us that if the click rates is the most important factor to make money. Yeah, I think they're intertwined, actually. Because the fun you explain that well, right. So they target a cheaper audience. And the best audience's clicks the most correct? Yeah. So in my head, these are interrelated. No. Don't see it. Yeah. Yeah. Yeah, I think, oh, everything together. Yeah.
112)	Speaker 2	Perfect. And with that, we conclude our interview. And I just have to say like, You are the last one, and you are the one that provided us. So many insights. So super happy.
113)	Participant 8	Wow. Thank you. That's a great compliment.
114)	speaker 1	No, thank you for participating. We learned a lot. And I'm actually like, kind of interested in like going and stalking your companies where it sounded real interesting.
115)	Participant 8	Yeah. And these questions made me think a lot and feel free to add me on LinkedIn. And also, I'm trying to bring people like you who do research academic research. I'm trying to connect them to my company, because we we were proud of being science driven. So once you're done with your thesis, and you have something that's publishable presentable, I would love to bring you to present if you're interested. Wow.
116)	speaker 1	Wow, thank you. Well, the pressure because
117)	Speaker 2	Thank you know that that's great. And I we have gotten so many links and key findings that of course, we will have to let you know, as soon as perfect we passed.
118)	Participant 8	Yeah, perfect. Let's let's stay connected on LinkedIn. And wishing you luck with your masters and you're doing something really cool. So,
119)	Speaker 2	so thank you, and you too.
120)	Participant 8	Enjoy your weekend. You to bye bye

Appendix 16: Can you ensure there is no bias in AI systems?

Participant	No bias in AI systems?	Comments from the participants
P1	No	Avoiding bias completely is hard, what matters is if the bias is intended or not.
P2	No	No way of ensuring it 100% but there are ways to make it better.
P3	No	Remarks that companies focus on profits and often don't care about bias.
P4	No	Participant argues that most companies have bias, it's more of a matter to be aware of it so one can decide how to deal with it
P5	No	You can never be 100% sure that there is no bias at all.
P6	No	Mentions that is really hard to ensure no bias in the systems, thus the final answer given was no
P7	N/A	N/A
P8	Depends on the system	Mentioned that in their systems they tried to ensure no bias, however it might not be possible in other systems. Putting an emphasis on the importance of designing models that can be explained and understood. (L46)

Appendix 17: Importance of diversity

Participant	Is diversity important to detect bias?	Comments from the participants
P1	It is important	Discuss that diversity provides a ‘’subsconscious understanding of things’’ opening people's minds that bias exist from different perspectives (l36)
P2	Yes definitely	Mentions that diversity plays a big role in their systems, Discusses how before there were many issues due to the lack of diversity in the engineering departments, reflecting that diversity helps in understanding things that are not so easy to see, specifically on the data. P2 proceeds to give an example on how products from women could be easily identified by a woman without having the gender attribute in the data. (L47)
P3	Its very important	Mentions that diversity is very important as it provides difference experiences to understand predictions about the collected data (L39)
P4	Yes but depends	Agrees on diversity being an important part of bias recognition, however the participants pays special attention on the kind of problem that it is being deal with, such as technical or related to humanity problems. Participants also mention that diversity does not automatically makes it better, however having diverse backgrounds might help. (L39)
P5	Yes Its very important	Note that a diverse team is very important as it provides different perspectives, that can help identifying issues in the data for example, (L44)
P6	Yes i love that	Answers that diversity is important, and their team is very diverse, which is something that P6 loves. L63
P7	Yes very important ‘’	Mentions that a diverse team is very important, specially when working with sensitive data, puts emphasis on the labeling of the data mentions that if a diverse team produces a similar result then that result might be good. Additionally P7 put emphasis on not just cultural and etchnic backgrounds but also academic backgrounds.
P8	Yes	Reflects on the importance of diverse teams as they help broaden peoples perspectives, adding that it does not only apply to soft science but also hard science, mentioning that both needs to be included in the teams (L65)

Appendix 18: Detailed overview of the participants thoughts regarding data bias

	Data bias	
	Data collection	**Data preparation**
P1	It's often the data that is to blame for bias (L9; L34). Open-source datasets often contain bias (L15).	Look at the data to see if anyone is singled out, overrepresented or if there are any proxy variables that can result in misleading results (L19). Further digging to find correlations (L21)
P2	Infrastructure that supports companies with data generation for certain data (L16).	As a data scientist it is hard to control the quality of the data that is given to you, as you are not part of the collection (L18). It is important to check the data pipeline and have assertions in your database. Sanity checks can also be used which is hard to do sometimes since it can be arbitrary meaning it's hard to check if the data is reasonable or not (L18). It's good to have people with knowledge in the domain cooperating with the people working with data to check the quality (L20). Making graphs can help finding outliers and better understand the data distribution (L26). Important to have data that is representative of the place the system is going to be used in (L41). Lack of representative data is the most common problem (L45).
P3	Data is produced by the customer, therefore the power of controlling the quality of collection is limited (L15).	In terms of quality checks, this is done by data engineers who analyze the data using different tools. They check if the data is reasonable or not and use different metrics to check if the data is true or not (L19). They are provided with raw data from their data engineering team that they need to aggregate (L23).
P4	Important to understand the problem one is trying to solve to know what data is needed (L17). It is hard to ignore the reality of the biases in the world if they are being added to	Puts an emphasis on making sure the data distribution matches the real world and is representative (L17). One needs to validate the data to see that what is shown in the dataset is actually an accurate representation of the world as we know it (L19). To check this and the data distribution it is important to understand the problem one is trying to solve (L19).

	the systems by the data collected (52)	When something goes wrong in the model it is good to go back to the data and analyze it to see if there is something wrong, missing or something that can be added (L19). How and if data cleaning should be done is dependent on the case and the problem that one is looking to solve (L28). Anomaly detection is also an important part and can be used to detect problems regarding imbalanced data (L35). When talking about Case 1 it is mentioned that the data is representing the real world and the real world is biased. It's a matter of historical data (L54).
P5	It's important that the generated data is relevant and representative enough for the case and the place it is going to be used in (L18).	One needs to make sure it is representative of the group it's going to be used for. Preprocessing is also important to make sure the data is good and not missing any important data points. If there are missing data points it's important to have a process for how to deal with it, will you delete it or will you try to find the right value, or are you going to proceed with an average? (L20).
P6	Data provided from client (L28).	Argues it's important to have representative and high data variety (L26). An iterative process where they keep asking the client for more data for a more representative dataset (L32). The participant defines high quality data as data with variety for the project the company is currently working on. Better to have a small dataset with high variety rather than a big dataset with low variety (L34). It's important that the data is balanced as well (L46).
P7	Do not collect more data than necessary (L16).	Data labeling is important and a stage where a lot of data bias could creep in (L16).
P8	Their company is interested in low quality data, therefore they don't have anything special in mind when collecting data (L25).	Important to have representative data. The data should also be balanced, and the dataset should represent the population you are aiming to use the system for (L29).

Appendix 19: Detailed overview of the participants thoughts regarding algorithmic bias

	Algorithmic Bias			
	Model development	**Model Evaluation**	**Model post processing**	**Model Deployment**
P1	There are often not malicious intents when building models (11) A lack of attention during the model process can intensify bias. (65) Representation bias can be combated during the development phase referring to Case 1 (47)	Overfitting models can lead to discrimination outcomes (21)	Not mentioned	There are certain places where machine learning can be dangerous (51)
P2	The necessary variables should be taken into account when building the model. P2 gives the example of measuring strength for grades in PE, if gender is not taken into account females would perform badly (30) When developing the process the 'historical' bias were not taken into account thus amplifying bias with the systems (68)	Understanding the output of the model is important to make sure is performing properly in different subgroups (39) Referring to Case 1 mentions the possibility the model was not evaluate in subgroups (58) Reiterates that model evaluation is tied with data collection, but data bias is more dangerous (58) Accuracy can be measure differently	Not mention	Not mentioned

		depending on the model development, if models are testing depending on accuracy it can help understand if a variable is needed or not (34) Model evaluation can be done with the wrong variables referring to case 1 (64)		
P3	Mentions is important to have bias free algorithms but do not specify where in the process, additional says that the most bias comes from data (35) Mentioning case 3 algorithms are being manipulated for a desired outcome which can affect the processing thus the outcome itself to become bias (56)	Links accuracy with profit, if a model is accurate does not mean that is fair (25)	Not mentioned	Not mentioned
P4	Mentioning case 3, is algorithmic bias, however, there is not a pinpointed stage on the process. Mentioning that in some cases depending on how the model is built there is not much that can be done in the data bias. Some problems are more about business value and the definition of the model objective (61).	If there is a lack of data, model evaluation can notice this, accuracy can then be tested in different classes, to make sure it performs well in important features (48) Is about thighing the whole thing together (48) Lets say that everything is properly done in the data preparation, then one can check on the model evaluation, to measure the quality	Not mentioned	P4 when referring to the compas case mentions that automatic decisions can lead to disadvantages to some parts of the population, thus systems can act unfairly (52) P4 touch upon a more complicate type of bias, that is not as easy to see as the ones in the data, for example when a same model is deploy in different locations to the ones they were originally intended to, mentions fraud detection

		of the steps in the data preparation and data collection. Iteration can be done in order to help remove bias in all parts (52) A high risk can involve data collection but also preparation and model evaluation (46)		mechanics, some countries might be more susceptible than others. (33)
P5	Algorithms will see if biases are identified in the data (42) Algorithm learned from a biases dataset thus is not able to recognize other features because it was trained in an specific way (58) referring to case 1 Referring to case 3 (68) model development as the model was build to act in an specific way causing bias (68)	Mentions that data collection and model evaluation go together (66) bias could have been fixed before entering the model but didn't happen thus it could have been seem in the model evaluation before deploying the model all together, Additionally to case 3, there were not any check to make sure the algorithm was favoring others (69)	Not mentioned	
P6	Choose model development as one can choose the attributes needed to use in a model (85) referring to case 2 It can be checked where the algorithms go wrong and if there are any missing clases, machines needs to be integrated over and over from the problem definition (36)	Data preparation and model evaluation for case 2 and 3 because somehow the preparation of the data can be check in the model evaluation Here it can be identified if the model is reading all parts needed or choosing to read specific p6 takes the example of an	Not mentioned	Model deployment regarding case 3 (101)

		animal clafisication to detect animals, however when tested they noticed it was testing the nose rather than the whole image, so all information from a label needs to be used. (61)		
P7	Referring to case 2 model development was the problem due to a fairness metric used in the system where one fairness metric favored one side rather than the other one, then it directly affected darked skin people (45)	P7 mentions that all stages of the model life cycle are subject to bias, from the ideation to the building of the model, testing and training the model with putting the model into production and even retiring the model. Thus P7 emphasize the importance of looking at the whole process thoroughly, (10)		A way of detecting bias can be outcome based, by mentioning is important to understand where a model is used, referring to using a model in an unpredictable and changing environment might lead to different outcomes. Takes the example of deduction of money laundry systems, even though the model and the data used are clean of bias, the models might flag people that are leaving war zones, making difficult for them to open bank accounts due to the unpredicted way their money is handled (21)thus it is important to understand not the inside of the model but also the outside and the way it is applied and used by people. Regarding case 3 p7 choose model deployment as is not just how the ad is programme but also by how it looks and how is tailored to certain demographics(53) (53)

P8	Mentions the importance of models being interpretable, as otherwise there can be black box models that one might not understand, thus harder to find where the problem lays(31) In order to ensure no bias in del model level, the mathematical libraries should be simple and interpretable, if the black box is incomprehensible, one must make sure its outcome then is revised, designated teams should be able to revise and make sure outcomes are not bias thus doing a manual review of the whole process , which can be expensive and timely (46) When talking about case 2 model development was mention as well as data because the data sidecitst could not explain the model itself, and that can be an introduction of bias (86)	Mentions that algorithms learn from data, if the data is bias, the model will be bias, however, when doing testing for example with adversarial testing, one could realize that the model that something was wrong (79) Referring to a famous case of amazon, if a test was created with specific datasets, the bias could have been spotted thus making the model evaluation also part of the problem (36) The way accuracy in a model is evaluated can lack diversity thus contributing to bias already found in the training sets (74) It is important to have human validation through all stages of the model, from the data feeding to the output of the model thus checking the quality in all phases (36)		Not mention

References

Angwin, J., Larson, J., Mattu, S. and Kirchner, L., 2016. Machine bias. In *Ethics of Data and Analytics* (pp. 254-264). Auerbach Publications. Available Online: https://www.taylorfrancis.com/chapters/edit/10.1201/9781003278290-37/machine-bias-julia-angwin-jeff-larson-surya-mattu-lauren-kirchner [Accessed 2 March 2022].

Baker, R.S. and Hawn, A., 2021. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, pp.1-41. Available Online:

Balayn, A. & Gürses, S. (2021). Beyond debiasing: regulating AI and its inequalities. European Digital Rights (EDRi). Available Online: https://edri.org/wp-content/uploads/2021/09/EDRi_Beyond-Debiasing-Report_Online.pdf?fbclid=IwAR3livF4Pp3zfHGsd3Nv068FjPF-sAHKwztHtvopb82_F6iMLU4NohX3rSis [Accessed 5 April 2022].

Balayn, A., Lofi, C. & Houben, G.-J. (2021). Managing Bias and Unfairness in Data for Decision Support: A Survey of Machine Learning and Data Engineering Approaches to Identify and Mitigate Bias and Unfairness within Data Management and Analytics Systems, The VLDB Journal, [e-journal] vol. 30, no. 5, pp.739–768, Available Online: https://doi.org/10.1007/s00778-021-00671-8 [Accessed 20 May 2022].

Barocas, S. & Selbst, A. D. (2016). Big Data's Disparate Impact Essay, *California Law Review*, [e-journal] vol. 104, no. 3, pp.671–732, Available Online: https://heinonline.org/HOL/P?h=hein.journals/calr104&i=695 [Accessed 11 April 2022].

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning. Available Online: fairmlbook.org. http://www.fairmlbook.org. [Accessed 1 Oct 2021].

Batini, C., Cappiello, C., Francalanci, C. & Maurino, A. (2009). Methodologies for Data Quality Assessment and Improvement, ACM Computing Surveys, [e-journal] vol. 41, no. 3, p.16:1-16:52, Available Online: https://doi.org/10.1145/1541880.1541883 [Accessed 5 April 2022].

BBC. (2015). Google apologises for Photos app's racist blunder. Available Online: https://www.bbc.com/news/technology-33347866 [Accessed 11 April 2022].

BBC. (2018). Amazon scrapped 'sexist AI' tool. Available Online: https://www.bbc.com/news/technology-45809919 [Accessed 13 April 2022].

Benbya, H., Pachidi, S., & Jarvenpaa, S.L. (2021). Special Issue Editorial: Artificial Intelligence in Organizations: Implications for Information Systems Research, Journal of the Association for Information Systems, vol. 22, iss. 2, pp.281-303. Available Online: https://aisel.aisnet.org/jais/vol22/iss2/10/ [Accessed 2 march 2022].

Bhattacherjee, A. (2012). Social Science Research: Principles, Methods, and Practices, 2nd edn, Tampa: A. Bhattacherjee.

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, Vol. 29, 2016, Curran Associates, Inc., Available Online: https://proceedings.neurips.cc/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html [Accessed 6 April 2022].

Britannica. (n.d.) Marvin Minsky - American scientist. Available Online: https://www.britannica.com/biography/Marvin-Lee-Minsky [Accessed 27 February 2022].

Bryman, A. (2016). Samhällsvetenskapliga metoder, Stockholm: Liber

Buolamwini, J. & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, in Proceedings of the 1st Conference on Fairness, Accountability and Transparency, Conference on Fairness, Accountability and Transparency, 21 January 2018, PMLR, pp.77–91, Available Online: https://proceedings.mlr.press/v81/buolamwini18a.html [Accessed 6 April 2022].

Caliskan, A., Bryson, J. J. & Narayanan, A. (2017). Semantics Derived Automatically from Language Corpora Contain Human-like Biases, Science, [e-journal] vol. 356, no. 6334, pp.183–186, Available Online: https://www.science.org/doi/10.1126/science.aal4230 [Accessed 6 April 2022].

Cambridge Dictionary. (2022). Bias. Available Online: https://dictionary.cambridge.org/dictionary/english/bias [Accessed 12 May 2022].

Chen, W. & Hirschheim, R. (2004). A Paradigmatic and Methodological Examination of Information Systems Research from 1991 to 2001, *Information Systems Journal*, [e-journal] vol. 14, no. 3, pp.197–235, Available Online: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365- 2575.2004.00173.x [Accessed 3 January 2022].

Cheng, X., Lin, X., Shen, X.-L., Zarifis, A. & Mou, J. (2022). The Dark Sides of AI, Electronic Markets, [e-journal] vol. 32, no. 1, pp.11–15, Available Online: https://doi.org/10.1007/s12525-022-00531-5 [Accessed 22 May 2022].

CNN. (2016). New Zealand passport robot thinks this Asian man's eyes are closed. Available Online: https://edition.cnn.com/2016/12/07/asia/new-zealand-passport-robot-asian-trnd/index.html [Accessed 22 April 2022].

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women, Reuters, Available Online: https://www.reuters.com/article/us-amazon-comjobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-biasagainst-women-idUSKCN1MK08G/ [Accessed 2 March 2022].

Dunphy, S. (2018). Women are seeing fewer STEM job ads than men: are marketing algorithms promoting gender bias? Available Online: https://www.europeanscientist.com/en/public/women-are-seeing-less-stem-job-ads-than-men-are-marketing-algorithms-promoting-gender-bias/ [Accessed 22 April 2022].

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M. and Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. nature,

542(7639), pp.115-118. Available Online: https://www.nature.com/articles/nature21056?spm=5176.100239.blogcont100708.20.u9mVh9 [Accessed 7 April 2022].

European Commission. (2019). Ethics Guidelines for Trustworthy AI | Shaping Europe's Digital Future, Available Online: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai [Accessed 11 April 2022].

Feller, A., Pierson, E., Corbett-Davies, S. & Goel, S. (2016). A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased against Blacks. It's Actually Not That Clear., p.5. Available Online: http://www.cs.yale.edu/homes/jf/Feller.pdf [Accessed 5 April 2022].

Fernandez, A., Garcia, S., Galar, M., Prati, R. C., Krawczyk, B. & Herrera, F. (2018). Learning from Imbalanced Data Sets | SpringerLink, Available Online: https://link.springer.com/book/10.1007/978-3-319-98074-4 [Accessed 7 April 2022].

Ferrer, X., Nuenen, T. van, Such, J. M., Coté, M. & Criado, N. (2021). Bias and Discrimination in AI: A Cross-Disciplinary Perspective, IEEE Technology and Society Magazine, vol. 40, no. 2, pp.72–80.

Fox, C., Levitin, A. & Redman, T. (1994). The Notion of Data and Its Quality Dimensions, Information Processing & Management, [e-journal] vol. 30, no. 1, pp.9–19, Available Online: https://www.sciencedirect.com/science/article/pii/0306457394900205 [Accessed 5 April 2022].

Frawley, W. J., Piatetsky-Shapiro, G. & Matheus, C. J. (1992). Knowledge Discovery in Databases: An Overview, 3, AI Magazine, [e-journal] vol. 13, no. 3, pp.57–57, Available Online: https://ojs.aaai.org/index.php/aimagazine/article/view/1011 [Accessed 31 March 2022].

Garcia, M. (2016). Racist in the Machine: The Disturbing Implications of Algorithmic Bias. World Policy Journal, 33(4), 111–117. Available Online: https://www.muse.jhu.edu/article/645268 [Accessed 30 March 2022].

García, S., Luengo, J. & Herrera, F. (2015). Data Preprocessing in Data Mining, 1st ed. 2015., [e-book] Springer International Publishing, Available Online: http://ludwig.lub.lu.se/login?url=https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=cat02271a&AN=atoz.ebs3191311e&site=eds-live&scope=site [Accessed 5 April 2022].

Ge, Y. (2019). A Survey on Big Data in the Age of Artificial Intelligence, in 2019 6th International Conference on Information, Cybernetics, and Computational Social Systems (ICCSS), 2019 6th International Conference on Information, Cybernetics, and Computational Social Systems (ICCSS), September 2019, pp.72–77.

Google. (n.d.). Responsible AI Practices, Google AI, Available Online: https://ai.google/responsibilities/responsible-ai-practices/ [Accessed 30 December 2021].

Grewal, D., Guha, A., Satornino, C. B. & Schweiger, E. B. (2021). Artificial Intelligence: The Light and the Darkness, Journal of Business Research, [e-journal] vol. 136, pp.229–

236, Available Online: https://www.sciencedirect.com/science/article/pii/S0148296321005294 [Accessed 22 May 2022].

Haenlein, M. & Kaplan, A. (2019). A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence, California Management Review, [e-journal] vol. 61, no. 4, pp.5–14, Available Online: https://doi.org/10.1177/0008125619864925 [Accessed 27 February 2022].

Han, J., Kamber, M. & Pei, J. (2012). Data Mining. Concepts and Techniques, 3rd ed., [e-book] Elsevier, Available Online: http://ludwig.lub.lu.se/login?url=https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=cat02271a&AN=atoz.ebs913105e&site=eds-live&scope=site [Accessed 5 April 2022].

Hankerson, D., Marshall, A.R., Booker, J., El Mimouni, H., Walker, I. and Rode, J.A., 2016, May. Does technology have race?. In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (pp. 473-486). Available Online: https://dl.acm.org/doi/pdf/10.1145/2851581.2892578?casa_token=9XvPoTyT9PkAAAAA:VuT5GFl4Miu9K2d2FTKomqZd0bz8OB1cHMWGsNYk4HxxHzbPw8hSXIx4HV8NWvQGvbZ8-iJwX2ujuC4 [Accessed 22 April 2022].

Hassani, B. K. (2020). Societal Bias Reinforcement through Machine Learning: A Credit Scoring Perspective, AI and Ethics, [e-journal] vol. 1, no. 3, pp.239–247, Available Online: https://doi.org/10.1007/s43681-020-00026-z [Accessed 27 April 2022].

Hellström, T., Dignum, V. & Bensch, S. (2020). Bias in Machine Learning -- What Is It Good For?, arXiv:2004.00686 [cs], [e-journal], Available Online: http://arxiv.org/abs/2004.00686 [Accessed 10 April 2022].

Hill, R. K. (2016). What an Algorithm Is, Philosophy & Technology, [e-journal] vol. 29, no. 1, pp.35–59, Available Online: http://link.springer.com/10.1007/s13347-014-0184-5 [Accessed 8 April 2022].

Hobson, S. & Dortch, A. (2019). IBM Policy Lab: Mitigating Bias in Artificial Intelligence, IBM Policy, Available Online: https://www.ibm.com/policy/mitigating-ai-bias/ [Accessed 6 January 2022].

Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M. & Wallach, H. (2019). Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?, in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, Maj 2019, New York, NY, USA: Association for Computing Machinery, pp.1–16, Available Online: https://doi.org/10.1145/3290605.3300830 [Accessed 19 April 2022].

https://link.springer.com/content/pdf/10.1007/s40593-021-00285-9.pdf [Accessed 23 May 2022].

Klein, H. K. & Myers, M. D. (1999). A Set of Principles for Conducting and Evaluating Interpretive Field Studies in Information Systems, *MIS Quarterly*, [e-journal] vol. 23, no. 1, pp.67–93, Available Online: https://www.jstor.org/stable/249410 [Accessed 3 January 2022].

Kordzadeh, N. & Ghasemaghaei, M. (2021). Algorithmic Bias: Review, Synthesis, and Future Research Directions, European Journal of Information Systems, [e-journal] vol. 0, no. 0, pp.1–22, Available Online: https://doi.org/10.1080/0960085X.2021.1927212 [Accessed 11 April 2022].

Lambrecht, A. & Tucker, C. E. (2018). Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads, SSRN Scholarly Paper, 2852260, Rochester, NY: Social Science Research Network, Available Online: https://papers.ssrn.com/abstract=2852260 [Accessed 20 May 2022].

Leavy, S. (2018). Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning, in Proceedings of the 1st International Workshop on Gender Equality in Software Engineering, ICSE '18: 40th International Conference on Software Engineering, Gothenburg Sweden, 28 May 2018, Gothenburg Sweden: ACM, pp.14–16, Available Online: https://dl.acm.org/doi/10.1145/3195570.3195580 [Accessed 25 February 2022].

LeCompte, M. & Goetz, J.P. (1982). Problems of Reliability and Validity in Ethnographic Research, Review of Educational Research, vol. 52, no. 1, pp. 31-60. Available Online: https://www.researchgate.net/publication/255615696_Problems_of_Reliability_and_Validi ty_in_Ethnographic_Research [Accessed 5 january 2022].

Lee, R. S. T. (2020). Artificial Intelligence in Daily Life. [Elektronisk Resurs], 1st ed. 2020., [e-book] Springer Singapore, Available Online: http://ludwig.lub.lu.se/login?url=https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=cat07147a&AN=lub.6535827&site=eds-live&scope=site [Accessed 29 March 2022].

Madgavkar, A. (2021). A conversation on artificial intelligence and gender bias. Available Online: https://www.mckinsey.com/featured-insights/asia-pacific/a-conversation-on-artificial-intelligence-and-gender-bias [Accessed 27 December 2021].

Marabelli, M., Newell, S. and Handunge, V., 2021. The lifecycle of algorithmic decision-making systems: Organizational choices and ethical challenges. The Journal of Strategic Information Systems, 30(3), p.101683. Available Online: https://reader.elsevier.com/reader/sd/pii/S0963868721000305?token=D351CFA511D11218C0106B337E5A1E6D28C65A7EAD23546ADDA1ACF41C600BC3DBB31E6B042C4F599D44947561BF9732&originRegion=eu-west-1&originCreation=20220428103545 [Accessed 25 April 2022].

McCulloch, W.S. and Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), pp.115-133.

McKenna, M (2019). Three notable examples of AI bias. AI business. Available at: https://ai-business.com/document.asp?doc_id=761095 [Accessed 2 April 2022].

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A., 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), pp.1-35. Available Online: https://dl.acm.org/doi/pdf/10.1145/3457607?casa_token=4aagReG3Rw0AAAAA:XdBcpzH8ogyv-

eerkWlRfN5mBgvdIeenkRjC9KSAO2xfZPQyhC-f4EZEfFG79sqc5cw6BX0lpf7sIQ [Accessed 10 April 2022].

Mero-Jaffe, I. (2011). 'Is That What I Said?' Interview Transcript Approval by Participants: An Aspect of Ethics in Qualitative Research, *International Journal of Qualitative Methods*, vol. 10.

Mikalef, P. and Gupta, M., 2021. Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Information & Management*, *58*(3), p.103434. Available Online: https://www.sciencedirect.com/science/article/pii/S0378720621000082 [Accessed 24 February 2022].

Mikalef, P., Conboy, K., Lundström, J. E. & Popovič, A. (2022). Thinking Responsibly about Responsible AI and 'the Dark Side' of AI, European Journal of Information Systems, [e-journal] vol. 0, no. 0, pp.1–12, Available Online: https://doi.org/10.1080/0960085X.2022.2026621 [Accessed 10 April 2022].

MIT News. (2016). Marvin Minsky, "father of artificial intelligence," dies at 88. Available Online: https://news.mit.edu/2016/marvin-minsky-obituary-0125. [Accessed 27 February 2022].

Mitchell, S., Potash, E., Barocas, S., D'Amour, A. & Lum, K. (2021). Algorithmic Fairness: Choices, Assumptions, and Definitions, Annual Review of Statistics and Its Application, [e-journal] vol. 8, no. 1, pp.141–163, Available Online: https://doi.org/10.1146/annurev-statistics-042720-125902 [Accessed 23 May 2022].

Mueller, J.P. and Massaron, L., 2021. Artificial intelligence for dummies. John Wiley & Sons.

Myers, M. D. (2013). Qualitative Research in Business & Management. 2nd edn, Thousand Oaks: Sage Publications Inc.

Myers, M.D. & Newman, M. (2007). The qualitative interview in IS research: Examining the craft. Information and organization, vol. 17, no. 1, pp. 2-26, Available online: [Accessed 17 April 2021]

Nasiripour, S., & Farrell, G. (2021). Goldman Cleared of Bias in New York Review of Apple Card. Available Online: https://www.bloomberg.com/news/articles/2021-03-23/goldman-didn-t-discriminate-with-apple-card-n-y-regulator-says [Accessed 4 December 2021].

Nedlund, E. (2019). Apple Card is accused of gender bias. Here's how that can happen. Available Online: https://edition.cnn.com/2019/11/12/business/apple-card-gender-bias/index.html [Accessed 29 December 2021].

Newell, S. and Marabelli, M., 2015. Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification'. The Journal of Strategic Information Systems, 24(1), pp.3-14. Available Online: https://reader.elsevier.com/reader/sd/pii/S0963868715000025?token=FD403041D3484E0ABB589AD42F4D6B36D51D74BABE03F0950D8187B517283776833369B2F011696AEC60C2F7704F11D3&originRegion=eu-west-1&originCreation=20220428014539 [Accessed 27 April 2022].

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E. and Kompatsiaris, I., 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), p.e1356. Available Online: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1356 [Accessed 10 April 2022].

Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. (2019). Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations, p.8. Available Online: https://www.science.org/doi/pdf/10.1126/science.aax2342?casa_token=XLCT1fncE-BIAAAAA:Ra1STfiDi3cycTJDDQmvxGiPm-CU_w6ag_o0yHIUS3xIaTP0pmWboS-kebNbk1JejNnktp4gvef73fPiX [Accessed 3 April 2022].

OECD. (2022). Artificial Intelligence. Available Online: https://www.oecd.org/digital/artificial-intelligence/ [Accessed 10 April 2022].

Olteanu, A., Castillo, C., Diaz, F. & Kıcıman, E. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries, Frontiers in Big Data, [e-journal] vol. 2, Available Online: https://www.readcube.com/articles/10.3389%2Ffdata.2019.00013 [Accessed 8 April 2022].

Orlikowski, W. J. & Baroudi, J. J. (1991). Studying Information Technology in Organizations: Research Approaches and Assumptions, *Information Systems Research*, [e-journal] vol. 2, no. 1, pp.1–28, Available Online: https://pubsonline.informs.org/doi/abs/10.1287/isre.2.1.1 [Accessed 6 January 2022].

Parikh, R.B., Teeple, S. and Navathe, A.S., 2019. Addressing bias in artificial intelligence in health care. *Jama*, *322*(24), pp.2377-2378. Available Online: https://jamanetwork.com/journals/jama/article-abstract/2756196?casa_token=4fNNUIDbtfIAAAAA:NxNUvcP2xTMAsA1jn3Le6LypFw1iOm-bAFV88qkOddI5Me1Vn0qNQGfeZ81rZTGHeeWdgMuYDsjY [Accessed 10 April 2022].

Patton, M.Q. (2015). Qualitative Research & Evaluation Methods, 4th edn, Thousand Oaks: Sage Publications Inc.

Pilkington, E. (2019). Digital dystopia: How algorithms punish the poor. The Guardian. Guardian Media Group. Available at https://www.theguardian.com/technology/2019/oct/14/automating-poverty-algorithms-punish-poor [Accessed 25 April 2022].

Recker, J. (2013): Scientific Research in Information Systems: A Beginner's Guide. Springer, Berlin Heidelberg.

Russell, S & Norvig, P. (2016). *Artificial Intelligence: A modern approach*. 4th edition. Pearson Education.

Russell, S. J. & Norvig, P. (2021). Artificial Intelligence : A Modern Approach, Fourth edition., [e-book] Pearson, Available Online: http://ludwig.lub.lu.se/login?url=https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=cat07147a&AN=lub.6806195&site=eds-live&scope=site [Accessed 29 March 2022].

Saxena, N., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D. & Liu, Y. (2018). How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness, [e-journal], Available Online: https://arxiv.org/abs/1811.03654v2 [Accessed 20 May 2022].

Spanakis, E. K. & Golden, S. H. (2013). Race/Ethnic Difference in Diabetes and Diabetic Complications, Current Diabetes Reports, [e-journal] vol. 13, no. 6, pp.814–823, Available Online: http://link.springer.com/10.1007/s11892-013-0421-9 [Accessed 11 April 2022].

Skiena, S. S. (2020). The Algorithm Design Manual, [e-book] Cham: Springer International Publishing, Available Online: http://link.springer.com/10.1007/978-3-030-54256-6 [Accessed 7 April 2022].

Sujay, L. (2022). Global Artificial Intelligence Market Revenues 2023, Statista, Available Online: https://www.statista.com/statistics/694638/worldwide-cognitive-and-artificial-intelligence-revenues/ [Accessed 16 April 2022].

Suresh, H. & Guttag, J. V. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle, Equity and Access in Algorithms, Mechanisms, and Optimization, [e-journal] pp.1–9, Available Online: http://arxiv.org/abs/1901.10002 [Accessed 6 April 2022].

Suresh, H. and Guttag, J.V., 2019. A framework for understanding unintended consequences of machine learning. arXiv preprint arXiv:1901.10002, 2. Available Online: https://courses.cs.duke.edu/spring20/compsci342/netid/readings/suresh-guttag-framework.pdf [Accessed 6 April 2022].

Telford, T. (2019). Apple Card algorithm sparks gender bias allegations against Goldman Sachs. Washington Post. Available Online: https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-allegations-against-goldman-sachs/ [Accessed 20 April 2022].

Tsaku, N.Z. and Kosaraju, S., 2019, April. Boosting Recommendation Systems through an Offline Machine Learning Evaluation Approach. In *Proceedings of the 2019 ACM Southeast Conference* (pp. 182-185). Available Online: https://dl.acm.org/doi/pdf/10.1145/3299815.3314454?casa_token=KkPfXXrtJh4AAAAA:_gugHZn3NwFQA08I7Q-Uk1oTE0RUgmSTeKHKqOLzae0c6idseg6jBi3rfKbs9yHTnAaXuroCR0g-gg [Accessed 7 April 2022].

UNESCO. (2019). First UNESCO recommendations to combat gender bias in applications using artificial intelligence. Available Online: https://en.unesco.org/news/first-unesco-recommendations- combat-gender-bias-applications-using-artificial-intelligence [Accessed 30 December 2021].

Walsham, G. (1995). The Emergence of Interpretivism in IS Research, *Information Systems Research*, [e-journal] vol. 6, no. 4, pp.376–394, Available Online: https://www.jstor.org/stable/23010981 [Accessed 4 January 2022].

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S.M., Richardson, R., Schultz, J. and Schwartz, O., 2018. AI now report 2018 (pp. 1-62). New York: AI Now Institute at New York University.

Zarya, V. 2018. The share of female ceos in the fortune 500 dropped by 25% in 2018. Fortune. Available Online: https://fortune.com/2018/05/21/women-fortune-500-2018/ [Accessed: 8 April 2022].

Zeng, L. & Wu, J. (2021). A Wo-Dimensional Research Framework for Analysing Dark Side of AI, in 2021 International Conference on Computer Engineering and Application (ICCEA), 2021.