Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

Fredrik Wilhelm Butler Wang

Approaches to Algorithmic Fairness: Empirical Study and a Sociotechnical Framework Proposal

Master's thesis in Informatics
Supervisor: Ilias O. Pappas
June 2023

NTNU
Norwegian University of
Science and Technology

Fredrik Wilhelm Butler Wang

Approaches to Algorithmic Fairness: Empirical Study and a Sociotechnical Framework Proposal

Master's thesis in Informatics
Supervisor: Ilias O. Pappas
June 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

Abstract

Machine learning (ML) systems have the potential to increase social differences and limit people's opportunities in life. These systems can be used in several sensitive environments and partake in life-changing decisions, and it is, therefore, essential to ensure that these decisions are free from discriminatory behavior that target certain groups or individuals. Discoveries of injustice and systematic discrimination caused by algorithmic outcomes have made headlines within numerous contexts. Researchers have become aware of the biases within these systems and significant literature on algorithmic fairness and bias has been proposed, attempting to address these issues.

This research is often concerned with mathematically defining fairness and findings methods for mitigating bias and unfairness within these definitions. However, if the contributions of the research community are to have a positive impact on the industry it is pivotal that these suggestions and solutions also have an understanding of real-world needs.

The research objective of this study is to improve the understanding of how companies advance toward algorithmic fairness, by identifying relevant factors through retrieving insight from practitioners that are concerned with algorithmic fairness. The author used semi-structured interviews as the data generation method in order to solve the research objective. By performing 9 semi-structured interviews, a systematic investigation of how practitioners approach and advance toward algorithmic fairness is conducted, identifying the techniques used and considerations taken along with observed challenges.

The study contributes to the area of algorithmic fairness by extending a sociotechnical view based on discoveries from analyzed interviews and the literature. The proposed contribution is a framework based on relevant factors that use a sociotechnical lens on algorithmic fairness.

This study reveals that practitioners are at an early stage in making considerations for algorithmic fairness. The results show that whilst practitioners are preoccupied with the literature, the social aspects and consequences are given more attention than statistical measures. This research highlights both general and case-specific aspects, revealing implications that practitioners must consider.

Sammendrag

Maskinlæringssystemer har potensialet til å øke sosiale forskjeller og begrense folks livsmuligheter. Disse systemene kan brukes i flere sensitive situasjoner og være involvert i livsavgjørende beslutninger. Det er derfor viktig å sikre at disse beslutningene ikke inneholder diskriminerende adferd rettet mot visse grupper eller enkeltpersoner. Oppdaging av urettferdighet og systematisk diskriminering forårsaket av algoritmiske resultater har fått mye oppmerksomhet i ulike sammenhenger. Forskere har blitt oppmerksomme på bias i slike systemer, og det er blitt publisert en betydelig mengde litteratur om algoritmisk rettferdighet og bias for å prøve å løse disse problemene.

Denne forskningen er ofte opptatt av å matematisk definere rettferdighet og finne metoder for å redusere bias og urettferdighet innenfor disse definisjonene. Imidlertid er det avgjørende at disse forslagene og løsningene også tar hensyn til virkelige behov dersom forskningsmiljøets bidrag skal ha en positiv innvirkning på bransjen.

Målet med denne studien er å forbedre forståelsen av hvordan selskaper jobber mot algoritmisk rettferdighet ved å identifisere relevante faktorer gjennom innsikt fra utøvere som er opptatt av algoritmisk rettferdighet. Forfatteren brukte halvstrukturerte intervjuer som datainnsamlingsmetode for å oppnå forskningsmålet. Ved å gjennomføre ni halvstrukturerte intervjuer, ble det gjort en systematisk undersøkelse av hvordan utøvere nærmer seg og jobber mot algoritmisk rettferdighet. Dette inkluderer identifisering av brukte teknikker, hensyn som tas og observerte utfordringer.

Studien bidrar til området for algoritmisk rettferdighet ved å utvide et sosioteknisk perspektiv basert på funn fra analyserte intervjuer og litteraturen. Det foreslåtte bidraget er et rammeverk basert på relevante faktorer som tar i bruk et sosioteknisk perspektiv på algoritmisk rettferdighet.

Denne studien viser at utøvere er på et tidlig stadium når det gjelder å vurdere algoritmisk rettferdighet. Resultatene viser at selv om utøverne er opptatt av litteraturen, får de sosiale aspektene og konsekvensene mer oppmerksomhet enn kun statistiske mål. Denne forskningen belyser både generelle og saksspesifikke aspekter og avdekker implikasjoner som må tas hensyn til.

Preface

I would like to thank Ilias O. Pappas for his supervision during the Master thesis process. I would also like to thank all the people who were interviewed and provided valuable insight for this project. Also a huge gratitude to those who otherwise supported this project.

Table of Contents

List of Figures

List of Tables

Abbreviations

List of all abbreviations in alphabetic order:

- **AI** Artificial Intelligence
- **GDPR** General Data Protection Regulation
- **IS** Information Systems
- **NTNU** Norwegian University of Science and Technology
- **RAI** Responsible Artificial Intelligence
- **RQ** Research Question
- **SMS** Systematic Mapping Study

Chapter 1

Introduction

The use of machine learning and algorithmic decision-making is as widespread as ever. Being applied in a range of applications and contexts, these systems and their outcomes have an effect on people's opportunities in life. At the same time, discoveries of unfair algorithmic outcomes that limit people's opportunities in life make headlines [15, 4, 3]. Algorithmic fairness has received increased attention from the research community, but mostly in terms of developing statistical definitions and mitigating biases in relation to these definitions [14]. Yet practitioners need to handle the many challenges of keeping a system fair, as bias can arise in several parts of the pipeline in an algorithmic decision-making system. More research should be done in order to understand how fairness, a social concept, is dealt with and modeled in technical systems, and address the real-world needs and challenges of practitioners.

The objective of this Master thesis is to *create a better understanding of how companies approach and advance toward algorithmic fairness, from initial ethical discussions to deployed solutions. More specifically to see what biases companies have to identify and mitigate, and identify factors, both technical and social, that are relevant and important to consider.* The findings from a multiple-case study with 9 participants from 8 different companies are presented. The implications of this research are the creation of the *The Extended Sociotechnical Framework for Algorithmic Fairness* and recommendations for future work.

The remainder of this chapter proceeds as follows: Section 1.1 presents the motivation for this thesis. Section 1.2 presents the research objective and the research questions. Section 1.3 specifies the boundaries of the research. Section 1.4 presents the research method that was chosen and the process followed. Lastly, Section 1.5 presents the outline for this thesis.

1.1 Motivation

Recent years have seen substantial advances in the field of Artificial Intelligence (AI). AI and algorithms affect many, arguably most, aspects of everyday life, such as hiring [48], loan approval [67], and more recently, after significant advancements, is being used in the generation of images [61], text [57] and code [32]. These AI algorithms are often thought to be objective and free of bias, which humans tend to have, however, this is not the case. Academics and regulators have found algorithms to reflect and exaggerate human and historical bias, and even introduce intricate biases themselves [51]. As a large amount of data is becoming more and more accessible, along with an increase in computational resources, AI algorithms grow more complex and less interpretable.

The term algorithmic fairness refers to technological solutions designed to prevent systematic advantages or disadvantages to certain subgroups in automated decision-making processes. From a purely technical standpoint, the goal of algorithmic fairness is to mathematically measure bias, then use this quantified result to minimize discriminatory practices in machine learning against specific groups. Fairness itself is not a technical concept, and thus algorithmic unfairness doesn't have to be solely viewed as a technical issue. Societal, organizational, and technical factors all play a part in the origins of unfairness in AI, and it could thus be approached from a sociotechnical standpoint [22].

Discoveries of algorithmic unfairness, and discrimination by algorithms have made headlines numerous times in a wide range of disciplines, grading algorithms containing bias that can limit later opportunities in life [4], unfair discrimination of minorities [3], and welfare benefit algorithms used by governments that discriminated on gender, age and language skills [15]. These serve as constant reminders that the use of AI comes with inherent risks.

Altogether this leads to algorithmic fairness being an intricate issue spanning multiple industries, with different contexts and starting points. This subsequently leads to an ecosystem with a lot of participants, such as researchers, policymakers, companies, and spokespersons, who all share a responsibility to overlook and administer the pursuit of fairness. However, with a diversified collection of stakeholders, there is yet to be a clear agreement on both defining the challenge and coming up with a proposed solution [74]. Especially prominent is the difference between theoretical research, or research on well-known fairness datasets, compared to corporate practitioners using unsanitized, real-world data where outcomes end up affecting people.

Algorithmic fairness has emerged as an advancing research field that has received a lot of attention. A Systematic Mapping Study (SMS) conducted in the fall of 2022 serves as the starting point for this thesis. The study "Algorithmic fairness: A systematic mapping study", which was the delivery for the course IT3915 - Master in Informatics, Preparatory Project, is included in Appendix A. The SMS identified several contributions by the research community, whilst also revealing a scattered literature with results that are difficult to directly compare due to the many definitions and metrics.

1.2 Research Questions

The research direction was selected based on a SMS conducted in the preparatory project in autumn 2022. Findings from the SMS provided interesting factors about new proposals and development. At the same time, what is perceived as fair is very context-dependent and thus how the contributions from the literature are applied could therefore vary tremendously. The results of this motivated the following research questions:

Research Question (RQ)1: How do companies approach and implement algorithmic fairness?

- **RQ1.1: What are the main challenges that companies face when implementing algorithmic fairness?**

RQ2: How do companies identify potential sources of bias in their algorithms, and what strategies are used to mitigate these biases?

RQ3: What are the factors that facilitate or hinder the implementation of algorithmic fairness?

1.3 Research Scope

In this research, the units of analysis are people whose work is concerned with algorithmic fairness. Algorithmic fairness is relevant in several disciplines and its importance spans wider than the field of computer science. In order to get a broad sense of how the subject was approached by practitioners, it was desirable to interview people in different types of companies as well as different roles.

Since algorithmic fairness is discussed within the use of machine learning or algorithmic decision-making, it's worth adding that to the author's knowledge, automatic decision-making is rarely fully automated (in Norway), especially in sensitive environments where fairness is of relevance, such as when the system handles personal data. There is usually some sort of human involvement, such as systems where only favorable outcomes for the end user are automatically approved, or outcomes made by algorithmic decision-making systems deemed unfavorable are reviewed by humans before a final decision is made. For instance, one can get an insurance claim automatically approved, but not automatically declined by a machine learning model. Another way of human involvement is when the system is an algorithmic decision support system [1, 38], and not a decision-making system. For instance, the output from a machine learning model is not the final decision but is given as input to a human who together with other available information makes the decision.

Based on this, interviewees were relevant for participation in the study if they worked in a company/organization that:

- Develops, maintains, consults, or audits algorithmic decision-making systems
- Implements or considers algorithmic fairness in the system

This resulted in a broad range of participants that had different backgrounds, roles, and experiences. This also meant that algorithmic fairness could be discussed in a societal, organizational, and technical context.

Data was collected through in-depth semi-structured interviews. All of the companies and organizations that interviewees worked for were Norwegian or had a Norwegian branch.

1.4 Research Process

In order to collect data, 9 semi-structured interviews were performed with people from 8 different companies and organizations. All interviews were transcribed and recorded, and all participants signed consent forms. Interviews were conducted over Microsoft Teams with video and audio, as participants weren't necessarily located in Trondheim. This form of process fitted both the time constraint of the project as well as the availability of the participants.

A thematic analysis approach by Oates was adapted to analyze the data [55]. The transcripts were analyzed and used to generate codes. These codes were grouped and used to create themes. These themes were further related to each other by viewing algorithmic fairness through a socio-technical lens, grouping the themes into two higher orders, social and technical. By building on the sociotechnical view of Sarker et al [65], the identified themes, combined with knowledge from the relevant literature, are used to build *The Extended Sociotechnical Framework for Algorithmic Fairness*, describing how algorithmic fairness is approached and perceived.

1.5 Thesis outline

The structure of the thesis is presented in this section. The thesis consists of 7 chapters. Chapter 1 presents the motivation behind the thesis along with the objective, scope, and process of the research. Chapter 2 presents the background theory necessary to follow subsequent findings and discussions. Chapter 3 presents the systematic mapping study that was undertaken in the fall of 2022 prior to this thesis. Chapter 4 presents the applied research method. Chapter 5 presents the results and themes from the multiple-case study, including quotes from the participants. Chapter 6 answers the research questions and presents the *The Extended Sociotechnical Framework for Algorithmic Fairness.* Chapter 7 concludes the paper by stating the implications of the proposed framework and makes suggestions for future work.

Chapter 2

Background

This chapter presents the theoretical background of this Master's Thesis. The aim is for the reader to have the necessary knowledge for understanding the theory and topics in subsequent discussions. Section 2.1 presents a general theory on machine learning, its techniques, challenges, and versions of systems that use machine learning in various ways with different degrees of human involvement, especially for systems where fairness is of relevance. Section 2.2 gives an overall introduction to the concept of fairness. Section 2.3 presents the sociotechnical perspective and how it views algorithmic fairness. Section 2.4 theory related to algorithmic fairness, including different definitions, metrics, and trade-offs. Section 2.5 presents relevant frameworks and toolkits used by practitioners. Section 2.6 clarifies the term bias, and explains related types and methods. Finally, Section 2.7 introduces a proposed AI Act by the EU that can have an effect on how practitioners think regarding the use of AI and thus affect the relevance of algorithmic fairness.

2.1 Machine Learning and Algorithmic Decision-making systems

This thesis revolves around algorithmic fairness through the use of machine learning and algorithmic decision-making systems. Machine learning is put shortly a field of study where computers are given the ability to learn from data [31].

A typical systematization of machine learning systems can be done by distinguishing between the amount of supervision that is given during training. The upcoming subsections describe the four major categories:

2.1.1 Supervised learning

In supervised learning, the training data also includes the solutions, called labels, in order to train algorithms that predict or classify outcomes [31]. Supervised learning is used to solve several real-world problems, and can do so at scale, examples include spam mail classification or arrival time predictions. In the spam filter example, it is trained on many emails along with the class, and it learns to classify new emails. Supervised learning algorithms include linear regression, decision trees, and neural networks.

This is the type of learning that will be most relevant to this thesis, especially classification tasks, as most literature on algorithmic fairness targets classification tasks [51].

2.1.2 Unsupervised learning

This approach used machine learning algorithms to analyze and cluster unlabeled datasets. The algorithms do this by discovering hidden patterns or groupings in the data without human intervention[31]. Examples of this type of learning technique include tagging and grouping customers based on their shared characteristics, or detecting anomalies. This technique is particularly useful when practitioners don't know what they are looking for in the data, as the algorithm can perform initial explorations in order to find hidden patterns and structures.

2.1.3 Semi-supervised learning

Algorithms that can handle partially labeled training data, typically a lot of unlabeled data and some labeled data, are called semi-supervised learning [31]. A good example of this in practice is image-hosting services, i.e. Google Photos or Photos by Apple, where it can group together photos of the same person through a clustering algorithm, and then the user labels each person, so one later can search for persons in the photo album.

2.1.4 Reinforcement learning

Reinforcement learning is similar to supervised learning, but instead of training on sample data, the models learn by trial and error. Successful outcomes reward the model and these outcomes will then be reinforced to develop an optimal recommendation or policy for the problem [31]. Typical use cases of reinforcement learning are models (often called agents) that learn to play games with large search spaces, such as checkers, or in the decision-making process of autonomous vehicles, where models are trained to take the best action in different driving situations to maximize safety and efficiency.

2.1.5 Challenges of Machine Learning

There are several challenges related to machine learning, including many that also apply to algorithmic fairness. This section presents some of these challenges.

Data Quality is a common challenge with machine learning, and missing data can impact the usefulness of a machine learning system. Low data quality can stem from missing, incomplete, inconsistent, inaccurate, duplicate, and dated data, and has been a problem ever since the early days of computing [34]. Even the most sophisticated machine learning models will struggle to make good predictions if the data it is given is of low quality. This concept is often called “Garbage in, garbage out”.

Overfitting and Underfitting are challenges related to the training of a model. If the model learns the training data too well and performs poorly on unseen data, it is overfitted. When the model is unable to properly capture the relationship between the input and output, it will result in a high error rate during both training and on unseen data, this is referred to as underfitting [31].

Model Interpretability revolves around understanding why a model made a certain outcome. The interpretability of a model is dependent on the algorithm's design and techniques like feature importance. Complex models like deep learning networks are referred to as "black boxes" as they don't provide explanations as to how they reached an output, as opposed to a decision tree [31]. For instance, a neural network that predicts who's pictured in an image might perform well in terms of accuracy, but understanding why it performs so well and what features it is utilizing to make a prediction is challenging. Understanding how a model came to a decision can be crucial in certain fields such as healthcare or finance.

Algorithm selection is a challenge that arises because there exists a vast amount of algorithms to select from, that each have their strengths and weaknesses. Selecting an algorithm that fits the task is a difficult process that could involve a lot of trial and error. Several techniques and methods for selecting and comparing models have been proposed, such as holdout methods using train and test sets to estimate performances [62].

2.1.6 Human-in-the-loop and Algorithmic decision support

When discussing machine learning and algorithmic decision-making, one uses the terms input and output in the sense that given some input a model will produce an output, and that this output is the final outcome that an end user will see. For instance, in a financial loan application, a user will enter relevant information that is transformed into inputs fed into the model, and the output of the model is either that the application is approved or declined. However a lot of deployed machine learning systems do not have this form of power, instead, humans are involved in the process to some degree, such as human-in-the-loop systems. When discussing the use of AI, a common but important misconception is that concerns of fairness only arise when automated decision-making, such as prediction systems, are present, one reason for this misconception is that much of the scientific work emerges from computer science [52]. But given that much of human decision-making also depends on predictions, many of the same issues apply, and straight-up rejecting automated decision-making is not equivalent to avoiding these problems. In fact, human design, modification, and interpretation are used in order to define fairness [73].

Two common methods of human involvement that are particularly relevant to the systems deployed by interviewees in this study will now be presented. The first is an automatic decision support system. Relevant to this thesis is a system where the output of a machine learning model is given as an input to a human, for instance, a case manager who will use this prediction as well as several other factors. It's worth noting that other versions of human-in-the-loop systems exist, such as human involvement in data labeling [76, 54]. Figure 2.1 shows a potential version of an automatic decision support system. The box called "Relevant information" in the figure could be the same data that has been encoded and fed into the model.

Another method is to have a human check the outcomes that have a negative effect on a user, for example, those instances where a model believes the insurance claim should be declined. Instead of automatically declining, an actual human also goes through the application as well before a final

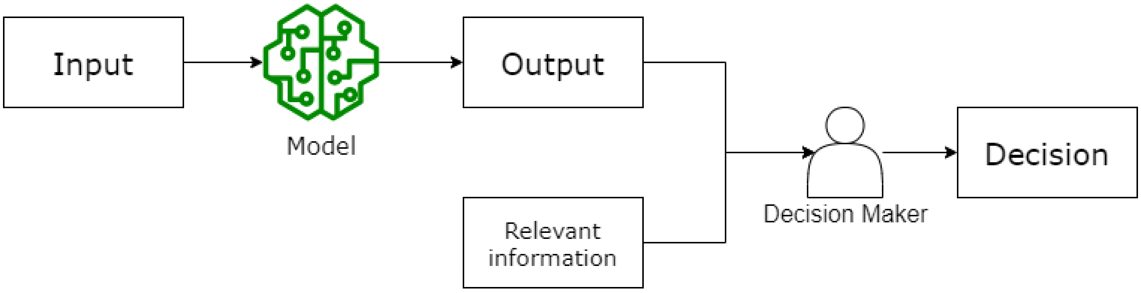


Figure 2.1: Potential version of an automatic decision support system.

outcome is given. However, if the model labels the input as positive, i.e. approves the claim, then the money is paid out immediately. Figure 2.2 visualizes a version of the no-automatic negative system where a machine learning classifier labels input as either positive or negative, where the negative outcomes are evaluated by a human before a final decision is made.

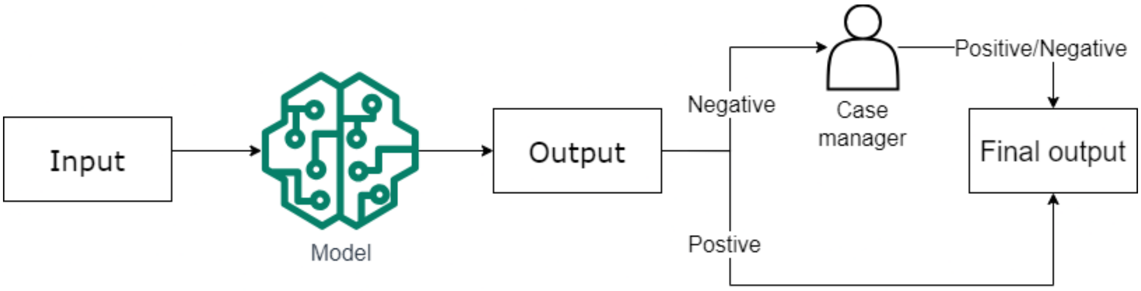


Figure 2.2: Potential version of a classification system with no automatic negatives.

2.2 Fairness

The concept of fairness is broad, and as a research field, it is quite complicated. There is also the case that there is no universal definition of fairness. Attempting to define literal fairness is also out of the scope of this thesis for several reasons. However, when talking about algorithmic fairness it is important to distinguish between the concept of fairness as a whole, and achieving algorithmic fairness. Section 2.4 will dive deeper into how the term algorithmic fairness will be used, while this section discusses fairness in general.

As mentioned, there is no universal definition of fairness, one reason for this is that culture and personal preferences affect how we perceive fairness. In society, achieving fairness is generally desired, but rarely fully achieved in practice. It is likewise difficult to say when fairness is actually achieved. In general, one could say that fairness involves taking a reasonable approach when dealing with things, and not taking sides. Another way of looking at fairness is the absence of discrimination. However both of these views are rather loose, and one could argue against it in several ways. For instance making the point that in some cases it is not possible to not take a side, or argue that the absence of discrimination could vary substantially depending on a nation's law for what counts as discrimination. As said the point of this thesis is not to figure out a way of defining fairness, instead, it is important to note that the term is complex and that fairness is not a static concept, but can change over time [6].

Fairness is predominantly envisioned as the principle of equal opportunity [6]. However, what equal opportunity means isn't necessarily agreed upon, and factors like what one seeks to achieve with equal opportunity play an important role. Other factors include understanding the causes of differences that occur between different groups, as well as how one would administer the cost of uplifting groups that historically have been at a disadvantage. The latter part is an important aspect when it comes to achieving fairness, as one thing is that one does not want to introduce discrimination, but since injustice has occurred throughout history, there will always be a matter of cost when it comes to repairing these injustices. Debates involving topics like reparations after historical injustice are particularly relevant to this aspect [21]

The concept of *unfairness* is typically expressed as limiting people's opportunity in life based on sensitive attributes, for instance, race or gender, or on attributes that are irrelevant based on the context. Examples of unfairness include gender as the reason for not being granted a loan, or ethnicity when performing a pretrial risk assessment, i.e. whether or not to release a person who is awaiting trial after being arrested. An observation regarding unfairness is that humans often agree that something is unfair, but often disagree on how to make something fair.

2.3 Sociotechnical Perspective

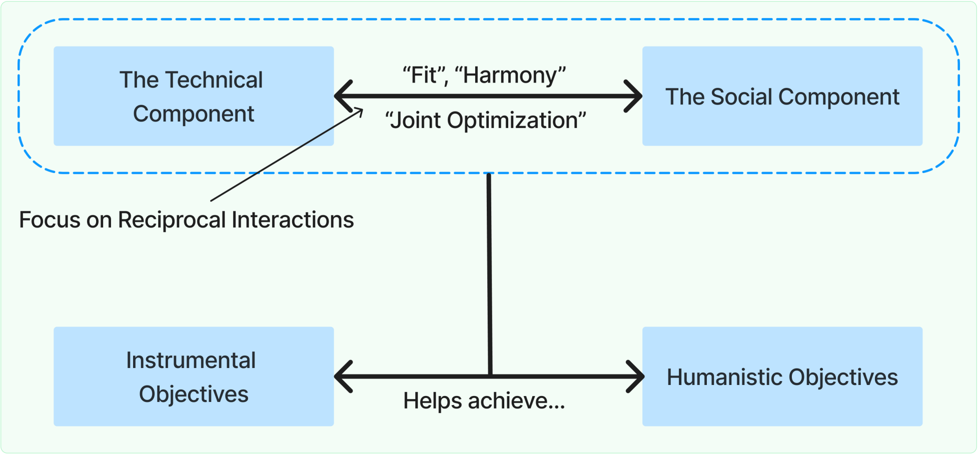


Figure 2.3: A Representation of the Sociotechnical Perspective in IS, adopted from [65]

The sociotechnical perspective in Information Systems (IS) is an approach that views technology and organizations as interdependent and intertwined systems. The sociotechnical perspective conceptualizes both the social and technical aspects as mutually interacting [65]. Opposed to looking at them as separate entities, the sociotechnical approach emphasizes that successful IS systems consider both aspects. Sarker et al. represent the sociotechnical perspective as two components, the technical component and the social component [65]. These components, through balance and reciprocal interactions, facilitate the achievement of both instrumental objectives and humanistic objectives. Figure 2.3 shows the essence of the sociotechnical perspective, adapted by Sarker et al. [65]. The focus on balance between components is pivotal, a change in the technical component, such as implementing AI to handle certain tasks, will affect the social system, such as people's way of working. When this balance is not considered, the implementation of the IS could lead to negative outcomes, such as resistance to use or ineffective use of technology. On the other side, if

IS is implemented successfully, both the technical requirements and the social implications should be understood. Implementing AI successfully requires not only technical expertise but also a deep understanding of the social context in which it operates.

In order to understand the role of algorithmic fairness and algorithmic decision-making in organizations, the sociotechnical perspective from Sarker et al. is adopted and applied for this research [65]. The following subsections explain the different parts of this perspective and how it has been applied.

2.3.1 Technical Component

The Technical Component refers to the technologies and tools used by the organization [65]. It involves factors like software, hardware, databases, networks, procedures, and other physical or digital resources. AI technologies can provide many services, such as task automation, insightful data analysis, or increase capabilities. However, implementing these technologies does not only consist of developing or buying the right services. Integration with existing systems, ensuring data quality, managing privacy, and managing the potential risks of the AI system. All these aspects must be carefully considered to ensure the effective use of AI.

2.3.2 Social Component

The Social Component involves people, groups, and their relationships within the organization [65]. It includes factors like organizational culture, work processes, politics, motivation, communication patterns, job satisfaction, team dynamics, skills, knowledge, and attitudes. The introduction of AI can have a significant impact on the people in an organization. Jobs may change or be eliminated, new skills may be required, decision-making processes may be altered, and organizational structures may need to be adapted. The sociotechnical perspective emphasizes that these social implications must be carefully managed. For example, management, communication, training, and support are all crucial to ensure that people understand and accept the changes brought by AI. Furthermore, ethical considerations like fairness, accountability, and transparency in AI decision-making are increasingly important.

2.3.3 Instrumental Objectives

Instrumental objectives refer to outcomes such as efficiency, productivity, and profitability [65]. These objectives look at how the technical system can be utilized to achieve specific goals of the organization such as improving accuracy, increasing speed, reducing costs, or enhancing capabilities. In the context of AI, an instrumental objective might be to automate a certain task, increase the accuracy of predictions, or enable faster decision-making, as AI offers the opportunity to process vast amount of information without fatigue or downtime.

2.3.4 Humanistic Objectives

Humanistic objectives refer to outcomes such as well-being, freedom, and quality [7]. They focus on the impacts that technology has on individuals and groups within the organizations. In the context of AI, humanistic objectives might involve designing systems that are transparent, understandable, fair, or that augment human capabilities rather than replacing people. Additionally, they can address ethical considerations, such as bias and privacy, in the design and use of AI systems.

Connecting Outcomes

Sarker et al. note that majorities of IS studies focused on instrumental objectives [65]. Yet there is also a need for humanistic outcomes. The sociotechnical perspective emphasizes that these two types of objectives are interrelated and need to be balanced. One justification for this is that the humanistic and instrumental outcomes can form a connected cycle that benefits both types of objectives, as the pursuit of humanistic outcomes generates positive actions, which then can lead to the creation of more positive instrumental outcomes [65]. Doing so can lead to positive reinforcement within the organization.

2.3.5 Sociotechnical Perspective in Other Studies

Other IS studies have utilized the sociotechnical perspective from Sarker et al. [65]. Dolata et al. see algorithmic fairness as a sociotechnical construct where a joint optimization between the Social and Technical component ultimately leads to improved instrumental and humanistic outputs, spanning beyond fairness in automatic decision-making systems [22]. Eulerich et al. adopt the sociotechnical perspective to develop and validate a framework for robotic process automation in audit tasks [26]. Kohn et al. apply the sociotechnical perspective in their study on remote work, utilizing the sociotechnical framework as a theoretical fundament for understanding dynamics in information systems [45].

2.4 Algorithmic fairness

Machine learning is applied to different issues, this means determining a universal definition of algorithmic fairness is difficult because what one would deem a fair outcome is context-dependent. As a consequence of this, several definitions of algorithmic fairness have been proposed in the literature, Verma et al. consider 20 definitions of fairness [70].

2.4.1 Definitions of algorithmic fairness

As stated by Dolata et al. algorithmic fairness is a sociotechnical phenomenon [22]. One reason behind this is that creating algorithms is a social practice, where algorithms can contain the assumptions and beliefs of their developer. In order to understand why an algorithm is biased one may also need more than simply the technical skills to dissect a model, but also knowledge about where the data originates from, how it was collected, and if it still reflects what the model is trying to predict. In a broader picture, algorithms are used in a wide range of applications, affecting millions of people. As a consequence of this, achieving fairness in the statistical sense is one thing, but in the end, the effects will be towards real people, who may not agree with the outcome because of the complexity of deeming something as fair or not. Multiple definitions of

Metric Name	Formula	Description
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Percentage of correct classifications
False Positive Rate (FPR)	$\frac{FP}{FP+TN}$	Measures the proportion of negative cases that are incorrectly classified as positive.
False Negative Rate (FNR)	$\frac{FN}{FN+TP}$	Measures the proportion of positive cases that are incorrectly classified as negative.
Precision	$\frac{TP}{TP+FP}$	Percentage of label 1 observations with the correct classification
Recall	$\frac{TP}{TP+FN}$	Proportion of true positives that were classified correctly

Table 2.1: Classification Metrics

algorithmic fairness therefore exist, and they will now be introduced, along with their respective groupings.

2.4.2 Metrics

most studies on algorithmic fairness concentrate on classification and typically binary classification problems [51]. This means that the classifier generates a prediction $\hat{y}$ based on an input x, which corresponds to an actual outcome y. So given a sample of several predictions, one way of measuring algorithmic fairness is by comparing the predictions $\hat{y}$ to the output y. It is typically assumed that x is a vector of features so that the dataset can be divided into different groups of attributes, such as ethnicity, gender, age, etc., as these variables can be used to determine the biases in an algorithm.

Fairness Metrics

As algorithmic fairness research has grown ever so popular, there has been an increase in statistical fairness metrics. These metrics are typically categorized as metrics for group fairness and metrics for individual fairness. There are different definitions for these types of fairness, although group fairness typically refers to the principle that certain groups should receive comparable treatment to the general population. This includes that the proportion of positive classifications within a group should be equivalent to the overall proportion of positive classifications in the population. For individual fairness the approach is similar, but from the perspective of individual members. Individual fairness is achieved when similar individuals receive equal treatment. What *similar* means depends on how it's defined for the considered scenario.

Table 2.1 shows performance metrics for classification problems, which serve as a base for metrics that measure algorithmic fairness. A non-exhaustive list of fairness metrics will now be presented. The notation used is similar to those of other works [35, 30]:

- $g_i \in \{G_1, G2\}$: The protected or sensitive attribute, G_1 means membership in the protected/unprivileged group, G_2 means membership in non-protected/ privileged group. This definition could be extended to account for more groups.

- x: All other attributes
- y: The actual outcome
- $\hat{y}$: The predicted outcome. $\hat{y} = 1$ indicates that the prediction label was 1. For definition purposes only, 1 is considered the 'favorable' label.

Group fairness

These definitions are regularly used for group fairness, although there exist several other metrics and definitions [51].

Statistical Parity

$$P(\hat{y} = 1|G_i) = P(\hat{y} = 1)$$

Statistical parity, also referred to as demographic parity, is based on the notion that group membership should not influence the likelihood of receiving a positive label classification [58]. It is one of the first fairness metrics proposed in the literature about fairness [44]. One can also measure the *statistical parity difference*, by taking the difference between the percentage of protected group members receiving classification 1 and unprotected group members receiving classification 1.

Disparate Impact

$$Disparate\ impact = \frac{Pr(\hat{y} = 1|G_1)}{Pr(\hat{y} = 1|G_2)}$$

Disparate impact is a similar metric to statistical parity but instead considers the probability of being predicted a favorable label based on group membership [58]. It measures the relative ratio between the two groups.

Equalized Odds, Equal Opportunity

$$FPR_{G1} = FPR_{G2}\ and\ TPR_{G1} = TPR_{G2}$$

Equalized Odds focus on the fairness of predictions across different groups. The principle asserts that a prediction model satisfies Equalized Odds if, for every group under consideration, the model's prediction accuracy is the same for both positive and negative outcomes. That is, among those who should have been predicted as positive (the true positives), the prediction is correct equally often across all groups (equal true positive rates). Similarly, among those who should have been predicted as negative (the true negatives), the prediction is correct equally often across all groups (equal true negative rates). This principle aims to ensure that the errors made by a prediction model are distributed fairly among all groups, avoiding any unintentional discrimination or bias. [36]

Fairness through unawareness This is an approach to fairness where an algorithm ignores sensitive attributes. The definition reads that an algorithm is fair as long as no sensitive attributes are used explicitly in the decision-making process [47] However, it does not guarantee fairness because other non-sensitive attributes can be correlated with sensitive ones, allowing the algorithm to indirectly discriminate. For instance, an algorithm that predicts job performance based on previous job titles might indirectly discriminate based on gender if men are more likely to have had certain job titles than women.

Individual fairness

As mentioned, individual fairness is based on the notion that similar individuals should be treated similarly. As with group fairness, individual fairness also requires access to the sensitive attribute. Additionally, it comes with some open-ended considerations to be made. One of them is to consider what attributes are relevant in regard to determining what makes individuals similar. Another consideration to be made is how to formalize nonquantitative attributes that remain relevant, for instance, addiction or emotional state. Designing a distance metric that measures the similarity of individuals would also require inputs from experts, and there is no guarantee that the definition will be unbiased due to inherent bias in the experts [70].

2.4.3 Trade-offs

When choosing the appropriate metric, it is important to take into account the relevant legal, ethical, and social context. As demonstrated previously, different metrics have distinct advantages and disadvantages. The primary trade-offs between various notions of fairness, as well as the fundamental trade-off between fairness and accuracy, will be highlighted in the subsequent paragraphs.

Metric trade-offs

Satisfying multiple notions of fairness simultaneously has been proven to not be possible in some cases [44]. In addition, achieving group fairness may come at the cost of individual fairness and vice versa. Statistical measures can also be insufficient, such as balancing error rates for all granulates of the data [14].

Fairness-accuracy trade-offs

This trade-off is heavily discussed in the literature, with the observation being that as fairness is achieved to a higher degree, there is a risk of compromising accuracy. This trade-off has been studied theoretically [40] and then later supported by empirical research [50].

In a real-world situation, whether or not this trade-off exists is context-dependent. For instance, if the system has some sort of legal obligation, such as non-discrimination, then sacrificing fairness for accuracy is not an option to consider at all. In systems where fairness is more of a desire than a requirement, then this trade-off is more relevant. In general, fairness-aware algorithms should aspire to achieve higher fairness without compromising accuracy, or other notions of utility for that sake.

Fairness-Privacy trade-offs

The pursuit of algorithmic fairness can inadvertently introduce new privacy risks, a concern that must be acknowledged in academic discourse and policy debates. To achieve fairness, machine learning algorithms often require access to sensitive attributes, such as race, gender, or the socio-economic status of an individual. While necessary for the mitigation of bias, the usage of such data can infringe upon the privacy rights of individuals, fairness and privacy are concerns that do not occur in isolation [12]. Moreover, even when these sensitive attributes are not directly used, techniques like disparate impact analysis or fairness through unawareness can infer them, potentially leading to indirect privacy violations. Furthermore, the legal landscape, including regulations like the General Data Protection Regulation (GDPR), may limit the extent to which sensitive attributes can be used, balancing the drive for algorithmic fairness against privacy requirements.

Some privacy-preserving approaches to algorithmic fairness exist, such as differential privacy in order to preserve individual privacy in a dataset [24], or federated learning where a model can be trained locally avoiding the need to exchange data through servers [49]. They allow for the development of more equitable algorithms while minimizing the exposure of sensitive data. However, these methods often involve a trade-off with the accuracy of the algorithm, highlighting the complex interplay between privacy, fairness, and performance in the design and application of machine learning models.

2.4.4 Considerations for fairness metrics

Many of these definitions share that they only consider 1 sensitive attribute at a time, however, in the real world it is possible for discrimination to take place at the intersection of two sensitive attributes. An example of this challenge is the analysis of failed employment discrimination lawsuits involving black women [16]. The famous analysis revealed that black women were unable to claim the discrimination as sexual discrimination because white women did not experience the same, nor could they make the case for racial discrimination because it did not apply to black men. Deeming if someone is being treated unfairly solely based on the distribution of each 'atomic' feature in the data is therefore not always enough, and intersections of features can reveal cases of unfairness.

Another important thing to keep in mind when measuring algorithmic fairness is that not all metrics are available at all times. Take the loan approval as an example, here only those predicted to repay their loan will be given a loan, meaning that all those who were not granted a loan are unobservable, making false negatives or true negatives impossible to measure.

2.5 Fairness toolkits and frameworks

Toolkits and frameworks for algorithmic fairness provide principles and guidelines aimed at ensuring that algorithms and automated decision-making systems do not discriminate against certain groups or perpetuate existing biases. These frameworks often include technical solutions, such as using diverse data sets and incorporating fairness metrics into algorithmic design. However, they can also emphasize the importance of considering the ethical and social implications of algorithmic decision-making, as well as promoting transparency and accountability in the development and deployment of these systems. Several of the large tech companies have developed their own toolkits, and some of them will now be presented.

2.5.1 Aequitas

Aequitas is an open-source bias audit toolkit. It provides an interface that both developers and policymakers can make use of in order to evaluate machine learning models based on several metrics [64]. It also includes a "fairness tree" which is designed to help practitioners select a relevant metric depending on the use case. This "tree" makes the practitioner consider their solution in several ways, such as what type of effect the algorithm has on individuals, and the capabilities of the algorithmic system.

2.5.2 Responsible AI Toolkit

The Responsible AI Toolkit is developed by Google/Tensorflow and includes a suite of tools that can be used to ensure responsible AI[1]. Fairness is one aspect that lies under the responsible AI paradigm, and the toolkit includes several tutorials, fairness indicators that can compute metrics and compare the performance of models for different subgroups, and tools to explore the impact of machine learning models.

2.5.3 AI Fairness 360

AI Fairness 360 is a framework developed by IBM [8]. The toolkit is open source and offers solutions for how to examine, report, and mitigate discrimination and bias [2]. Several bias-mitigation algorithms are offered, as well as metrics that measure both group and individual fairness. The framework offers easy access to several fairness datasets such as The Bank Marketing dataset, The COMPAS dataset, The Adult dataset, and the German Credit dataset. It also includes an interactive demo to visualize and compare bias and bias-mitigation techniques.

2.6 Bias

A human being makes decisions based on a lot of factors, some factors are obvious, and some we may not be aware of. As humans the decisions we make may also change as we become tired or agitated. Humans tend to have preconceived or unreasoned opinions of either someone or something, and these opinions are often prejudicial, close-minded, or unfair. We call these tendencies bias, human bias. Human biases have the possibility to affect decisions taken in a wide range of domains and sectors.

In machine learning, however, bias can be considered an overloaded word, as it can refer to several things. For instance, the “bias term”, also often called the intercept term, is a constant used by a model to compute a prediction [31]. The term is also used when referring to both the bias in the algorithm, and it may refer to statistical bias which affects the accuracy of a model’s outcomes. And so one may state that a model that predicts arrival times for buses is *biased* if it consistently

[1] https://www.tensorflow.org/responsible_ai

[2] http://aif360.mybluemix.net/

predicts that a bus will arrive earlier than it actually does [23]. While both of these interpretations of bias affect the algorithm, in algorithmic fairness and fair machine learning literature, the term bias is usually meant to describe when the outcomes are systematically disadvantageous for an unprivileged group, in the following discussions about bias, this is the intended meaning of the word.

2.6.1 Types of bias

There are several causes that could lead to unfairness in algorithmic decision-making systems. Typically these causes are referred to as biases, the following is a non-exhaustive list of potential biases:

- **Dataset bias** is the bias that exists in the data used for learning. This stems from biased measurements, historical bias from human decisions, or errors and inaccuracies in reports. An algorithm is only as good as the data it's trained on, so when there exists a bias in the dataset, the algorithm will in essence replicate this bias [51].
- **Historical bias** is the bias from socio-technical issues in the world that is already existing and can be seen as a subgroup of dataset bias. Even though the sampling and feature selection is perfect, the data may still contain historical bias [51].
- **Bias from proxies** is when certain attributes can act as proxies for sensitive attributes, and therefore differentiate privileged and unprivileged groups. Attributes that act as proxies are non-sensitive but can be used to derive sensitive attributes. When a dataset contains attributes that act as proxies, machine learning algorithms may make decisions based on sensitive attributes while appearing to only use acceptable attributes, such as triangulating race from other features [29], using salary as a proxy for gender, or finding someone's religion based on family structure. This type of bias can also be seen as a subgroup of dataset bias.
- **Algoroithmic bias** is when the input data is unbiased, and it is the algorithm itself that adds the bias, such as from design choices in the algorithm [5]. For instance when the algorithm's objective is to have a low average error, such as minimizing the *Mean Absolute Error*, then the predictions could benefit the majority group over minorities.
- **Feedback loop** is a phenomenon that occurs when there is bias in the data, which can lead to biased algorithmic outcomes, which leads to biased human actions. Then these new actions become training data, which can bolster and even increase the existing sources of bias [51].
- **Human bias** is bias that affects the actions of humans. One version of this is social bias where other's actions affect our own actions [5].

There are several methods that target bias in algorithms, but in general, they fall under three categories:

- **Pre-processing**. This technique tries to modify the data in order to remove underlying discrimination, such as by improving the demography of the data. A requirement for this technique is that the algorithm must be allowed to modify the data [19].

- **In-processing.** In-processing techniques try to remove discrimination during the training process of the machine learning model. To change the learning process is not always allowed, but when it is then this technique can be applied, typically by either changing the objective function for the model or by adding a constraint to the model [9].

- **Post-processing.** These techniques are applied after the training step. When it's not possible to modify the training data or alter the learning algorithm then these techniques are suitable. Typically a function would be applied that alters the model's predictions in order to correct or mitigate bias and unfairness [8].

Figure 2.4 extends on the automatic decision support systems from Figure 2.1, showing where different types of biases can affect different parts of a system. The effect of these biases can also be extended to affect the model during training, as well as the *Relevant information* to be biased as well.

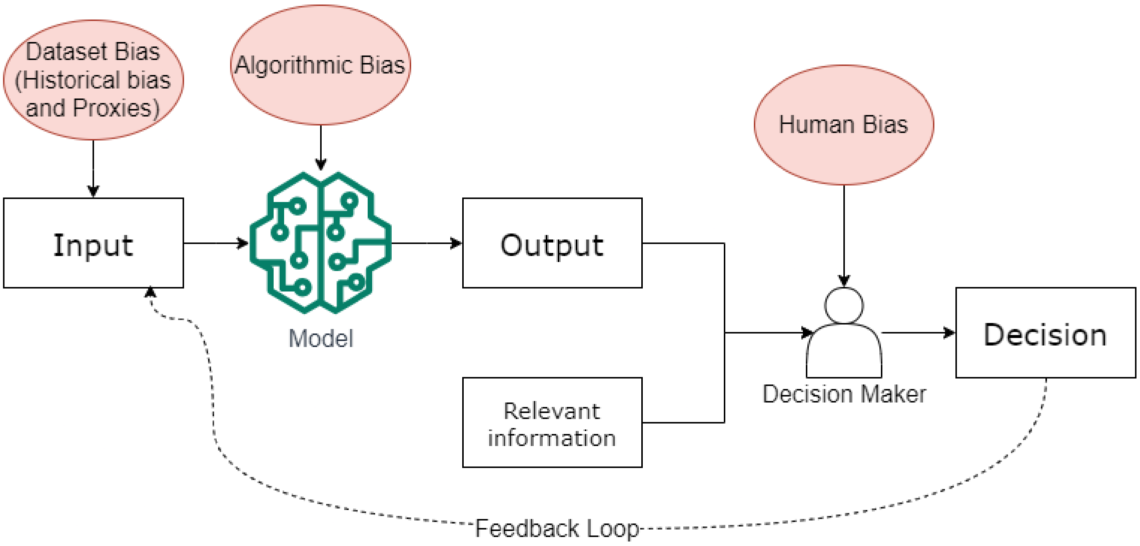


Figure 2.4: Types of biases that can arise in an automatic decision support system

2.7 European Union AI Act

The European AI Act is a proposed legislation from the EU [69]. The act will govern all who provide a product or service using AI. It is mainly aimed at AI usage in the public sector and law enforcement, but will also affect all AI companies. The Act proposes to classify different AI tools by their risk level. The law will not replace but overlap with existing laws such as General Data Protection Regulation (GDPR) [68], with the new proposal being more expansive as it is not restricted to personal data. The new proposal will affect the use of AI in Norway, and the Norwegian government has also issued a statement on the proposal [33]. Although the act is not final it is mentioned because it is something that practitioners have to prepare for and consider in both existing and new systems.

AI is split into 4 different levels of risks by the Act, *Minimal* (spam filters, video games), *Limited* (chatbots, deep fakes), *High* (Education, employment, law, justice, immigration) and *Unacceptable* (social scoring, facial recognition, manipulation) Risk. There are four different levels, that is based on the intended use of the system, and since there is no sliding scale, in theory only the *High Risk* category is subject to heavy regulation, with only small obligations for *Limited Risk* systems. In descending order, the different risk levels are presented:

Unacceptable Risk. These AI services are considered unacceptable, for instance, because they violate fundamental rights and therefore contravene with the EU's values. Some services considered unacceptable are: Subliminal techniques, i.e. AI that could distort people's behaviour and cause harm. Manipulation, such as exploiting vulnerabilities due to people's age or mental ability. Social scoring: The use of a social behavior system, typically used or created by a public authority to evaluate how trustworthy people are.

High Risk. These systems are considered High risk and must comply with extensive regulations, but are not banned from usage. There are some essential requirements that systems in this category must follow, such as data governance, technical documentation, transparency and providing users with information, record keeping, human oversight, robustness, accuracy, and security. As these systems become more and more widely used, the bigger the worry for discrimination and bias by these systems.

Relevant sectors where these systems play an important role include critical infrastructures such as transport, where the use of AI could put the life and health of citizens at risk. In education, these systems affect access to education or the course of an individual's professional career, such as automated grading of exams. In regard to employment, it affects access to self-employment, such as automated hiring or software that automatically processes CVs. Essential private and public services, such as automated credit scoring systems in the private systems, or automated systems for welfare benefits are also highly relevant for this category.

Limited Risk. AI systems in this category must provide transparency and disclosure, and regulations here overlap with current GDPR requirements, such as being transparent about personal data processing. Systems in this category include chatbots and systems that generate deep fakes or synthetic content.

Minimal Risk. The act proposes that these AI systems' main regulations should be voluntary codes of conduct. Instances in this risk category include spam filters or video games using AI.

If a company is found breaching the proposed AI Act a fine of up to 30 million euros or 6% of global profits (whichever is higher) can be issued [25]. A possible critique of the act is its inflexibility, so given that a completely new AI system should appear, or a system is used in an unforeseen way, it is difficult to label it in a higher-risk category.

In terms of algorithmic fairness and bias, rules about datasets used for training are introduced, with concerns for error and discrimination generated by partial, erroneous, or historically biased data. The proposal also suggests that systems should be designed in a way that they can be overseen by humans, which includes allowing a person to spot anomalies and biases [25]. The EU AI Act can therefore play an important role in how Norwegian companies implement fairness in their AI systems.

Chapter 3

Systematic Mapping Study

During the fall semester of 2022, a Systematic Mapping Study (SMS) was conducted. This chapter presents a summary of the study "Algorithmic fairness: A systematic mapping study" whereas the full study is available in Appendix A. The study was conducted in order to gain an updated view of algorithmic fairness and understand what contributions recent research produce. Through a developed classification schema, a mapping of 136 papers from 2018 to 2022 was conducted. Results revealed an increase in publications each year, with a shift toward technical and empirical research. Frameworks were found to be the most popular output in recent research, and with classification problems being where algorithmic fairness was considered the most. The results aligned with similar work such as Dolata et al. [22] and the work of Kordzadeh and Ghasemaghaei [46].

The rest of the chapter proceeds as follows: Section 3.1 explains the research process, Section 3.2 presents the main findings and synthesized results, Section 3.3 concludes the study and presents suggestions for future work.

3.1 Research Method

A systematic mapping study consists of identifying, categorizing, and analyzing existing literature relevant to a certain topic [60]. The aim of a mapping study is to get a comprehensive overview of a research topic and use this to attain an assessment of current literature [43]. In addition, the study can reveal research gaps or be used to make suggestions for future research [43]. Following a process influenced by Kitchenham's procedures [42], research questions were identified, a search strategy was developed and data was extracted through the study selection.

3.1.1 Research Questions

Algorithmic fairness has been viewed as an emerging field, that is going through a research "boom" [22]. Conferences such as ACM Conference on Fairness, Accountability, and Transparency (FAccT), and workshops like European Workshop on Algorithmic Fairness (EWAF) have received increased attention. The purpose of this study is to investigate the present condition of research in the field of algorithmic fairness. The goal of this mapping study is to offer a refreshed perspective on algorithmic fairness research and discover emerging research trends. The research objective motivated the following research questions:

- RQ1: How has research on Algorithmic Fairness changed between 2018 and 2022?
- RQ2: How do technical frameworks achieve Algorithmic Fairness?

The paper presents results from research on algorithmic fairness between 2018-2022. The study follows a similar search strategy from state-of-the-art reviews [22], resulting in 136 relevant articles. The results were classified in order to answer the research questions. In order to address RQ1 relevant papers were addressed based on their context, technical level, type of research, fairness focus, and output. For RQ2, papers whose output was a type of framework were further classified to understand what focus area they implement fairness in. The analysis reveals a shift in research focuses within the field of algorithmic fairness.

3.1.2 Data Sources and Search Strategy

During the initial stage, searches on Scopus were conducted to gauge the variety of outcomes that could be obtained from different search strings. This step was essential in identifying which keywords would yield the most relevant search results. The works of Dolata et al. offered a systematic review of 310 articles about algorithmic fairness, spanning from 2017 to 2020 [22]. In relation to RQ1, it could be beneficial to replicate this search but extend the timeframe to cover 2018-2022. This would help understand how the landscape of algorithmic fairness research has evolved since the study by Dolata et al. concluded [22].

The final search string is the same as the one used in Dolata et al. [22], which is an extensive combination of several terms designed to capture a broad range of relevant research.

> ((”*fair*” PRE/1 (”ML” OR ”machine learning” OR ”AI” OR ”artificial intelligence”)) OR ((”algorithmic*” OR ”AI” OR ”ML” OR ”machine learning” OR ”artificial intelligence”) PRE/1 (”fair*” OR ”justi*” OR ”bias*” OR ”unfair*”)))

3.1.3 Study Selection

The study selection is illustrated in Figure 3.1, with each step and the number of papers involved at each step. Searching Scopus with the search string resulted in 973 papers. Papers were then excluded based on publication between 2018 and 2022, subject areas of Computer Science and Business, Management and Accounting, document types of Articles and Conference papers, source type as Journals, as well as only papers written in English.

After restricting the search results as detailed previously, each paper’s title and abstract were thoroughly read and categorized as either relevant or not. To facilitate this process, a set of criteria for inclusion and exclusion was formulated. The final criteria adopted for this purpose are as follows:

Inclusion:

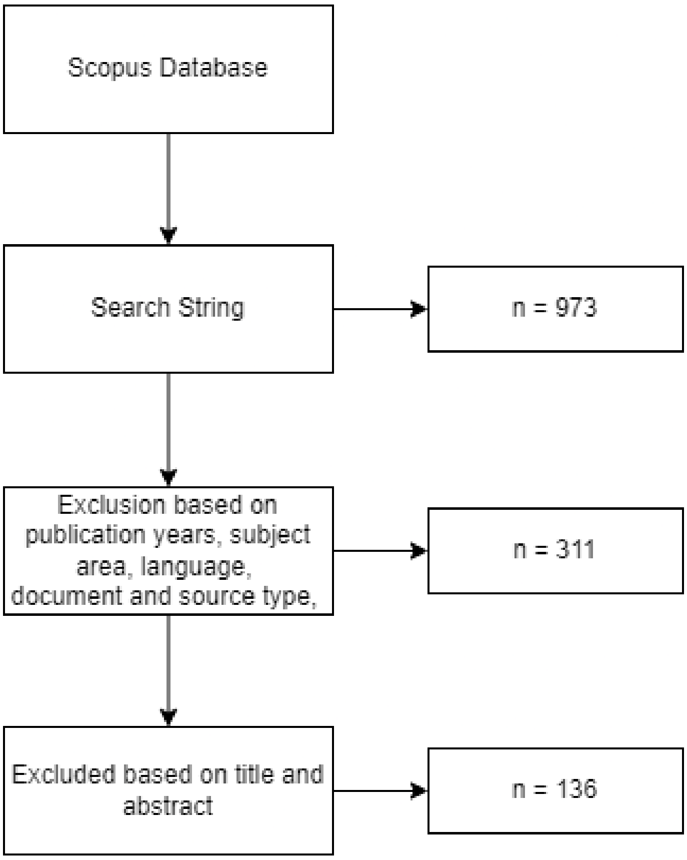

Scopus Database
Search String
n = 973
Exclusion based on publication years, subject area, language, document and source type,
n = 311
Excluded based on title and abstract
n = 136

Figure 3.1: The study selection process

1. Discusses fairness related to a technical solution
2. Makes a contribution to fairness or discusses fairness as a core concept

Exclusion:

1. No individual contribution (editorials, commentaries, calls for papers, or tutorials)
2. Words in query not used in the intended meaning
3. Refers to systematic deviation and not actual unfair treatment
4. Only refers to unfairness in general terms, no link between technology and discrimination
5. Only refers to fairness in the future work section or as a motivation

3.1.4 Threats to Validity

For this project, Scopus was the sole platform used to carry out all the research. This approach, while streamlined and convenient, may have limited the breadth of the search results, and potentially left some relevant academic papers undiscovered.

Scopus, as a reliable and comprehensive abstract and citation database of peer-reviewed literature, served us well in most respects. However, solely relying on it might create certain limitations. There is a multitude of other databases and search engines that could provide a wider array of

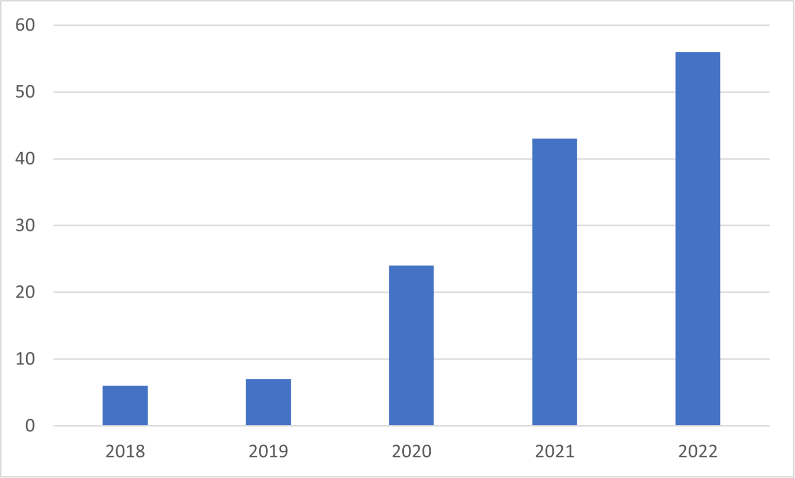


Figure 3.2: Publication frequency, 2018-2022.

relevant studies, and each has its unique strengths and focus areas. Using other search engines or screening conferences such as ACM Conference on Fairness, Accountability, and Transparency (FAccT). Moreover, the deployment of manual search methods could also have been beneficial. Although time-consuming, manual searches, such as directly searching the websites of relevant journals or referencing the bibliography of key papers, can yield more comprehensive results. They often uncover papers that automated search engines might miss due to the limitations of their algorithms. Thus, in retrospect, it's clear that broadening the search strategy by integrating another search engine or conducting manual searches, could have enhanced the robustness of the research and reduced the chances of missing out on any significant, relevant papers.

Bias from personal opinions is a threat to validity, especially when there is no co-author to assess selections. Following the recommendations from Kitchenham, the inclusion and exclusion criteria were decided before the study selection started [42].

3.2 Synthesized Results

This section presents a summary of the results in regard to the research questions defined in Section 3.1.1, the full results including the complete classification are available in Appendix A.

3.2.1 RQ1: How has research on Algorithmic Fairness changed between 2018 and 2022?

Figure 3.2 shows the publication frequency of papers related to algorithmic fairness. There is a clear increase in studies, implying a rapidly growing field. This increase agrees with what others reviews on fairness report [59, 11].

Papers were classified in regard to the research being conceptual or empirical. Figure 3.3 shows the distribution based on the papers being technical or non-technical. It's observed in particular that empirical research is most common for algorithmic fairness and that it is mostly technical.

Most research was found to be conducted in a generic manner, yet it was found that an increase

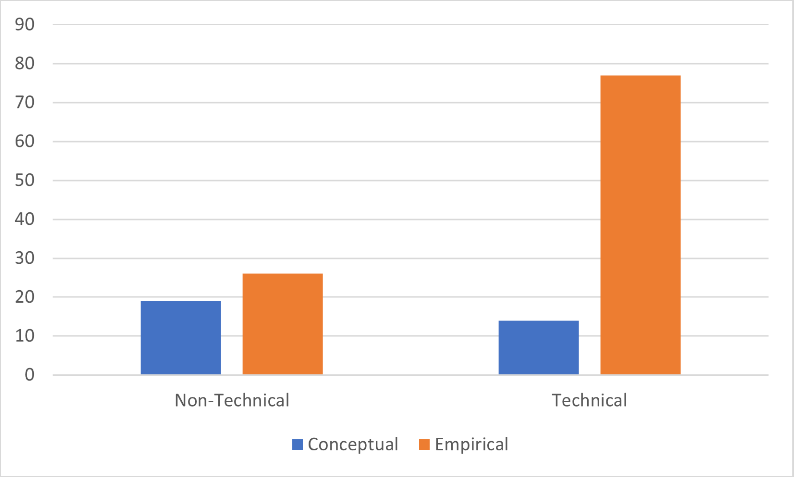


Figure 3.3: Number of technical and non-technical papers given type of research.

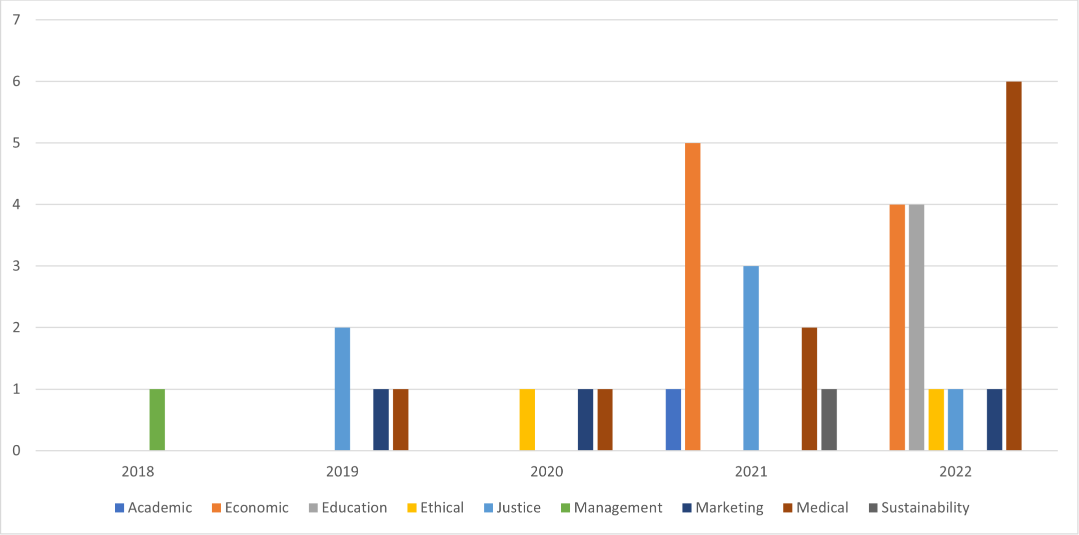


Figure 3.4: Domain-specific research, 2018-2022.

in overall literature has also led to an increase in domain-specific research. Figure 3.4 shows that the economic and medical fields receive the most attention.

3.2.2 RQ2: How do technical frameworks achieve Algorithmic Fairness?

All the relevant papers that proposed a framework were further classified based on their focus area for combating algorithmic unfairness. All the different focus areas were noted, and Table 3.1 shows the distribution of focus areas. The results give a clear indication that the proposed frameworks from the literature are mostly concerned with classification problems, which include both binary and multiclass classification.

Focus Area	Frequency
Adversarial learning	3
Auto-encoding	1
Casual inference	2
Classification	35
Clustering	1
Language model	1
Non-specific	4
Recommender systems	3
Regression	2
Word embedding	1

Table 3.1: Focus areas of technical frameworks that combat algorithmic bias and unfairness

3.3 Conclusion

A systematic mapping method has been applied to analyze algorithmic fairness literature. A total of 136 papers were extracted and classified. The study provides an extensive view of algorithmic fairness literature, in an attempt to understand the research focus and trends between 2018-2022.

Regarding RQ1, it was discovered that papers tended to be more technical, which could stem from the inherently technical nature of algorithms. Other studies also found that most studies were technical [22]. Empirical research was also more common than conceptual research, and one reason behind this could be that empirical research has more concrete and presentable results that could be more attractive to researchers and journals. Most literature was found to be conducted in a generic context, although there is an increase in domain-specific research within the economic and medical fields. It was also observed that several studies use popular yet old and criticized datasets that might not reflect the data intended or expected for the systems and that these studies instead should attempt to follow guides and updated overviews of fairness datasets, such as the work of Fabris et al. [27].

Regarding RQ2, frameworks were found to be the most common contribution from technical research on algorithmic fairness. It was further discovered that classification tasks are the most common focus area. This was also found by other studies [51]. A possible explanation for this is that classification is a broad topic and that the outputs of these algorithms are what most fairness metrics are designed to measure, for instance, false negative rates.

Future work can focus on possible solutions to how frameworks and metrics for algorithmic fairness can be more transferable between research and industry use. Work that focuses on applying algorithmic fairness research in real-world systems could be beneficial. Research and development of frameworks that accommodate multiple fairness definitions could help alleviate concerns over compatibility and facilitate easier evaluation. Future research could also focus on understanding how algorithmic fairness is understood depending on the context and industries, and identify unique practices or challenges.

Chapter 4

Research Method

The presentation of the research method consists of the methods and strategies employed in the process throughout this Master's thesis. The chapter aims to present how the research in this Master's thesis has been conducted and presents a detailed view of the series of events, and the reasoning behind them, that ultimately leads up to the results in chapter 5 and the discussion in chapter 6.

For this Master's thesis a strategy adapted from what Oates refers to as case studies is employed [55]. According to Oates, a case study centers its focus on a specific instance of the subject under investigation, whether it be an organization, a department, an information system, a development project, a decision, or similar entities. This particular case is thoroughly examined, employing a range of data generation methods, such as interviews, with the objective of attaining a comprehensive and detailed understanding of the case's intricacies, relationships, and processes [55]. The aim is to gain profound insights into the dynamics and complexities inherent in the case being studied. A case study has several characteristics, particularly relevant to this study is the "Focus on depth rather than width", and the "Natural setting". The first characteristic is rather self-explanatory as the goal is to obtain detailed information about the topic that is how the approach and implementation of algorithmic fairness is done by practitioners in different industries. The latter refers to the case being examined as pre-existent, and taking effect in a real-world situation rather than a controlled, artificial environment. The researcher steps into an already occurring case, and it normally continues to exist after the research is conducted, hopefully with as little interruption as possible [55]. Interviews were chosen as the method for data generation. To acknowledge the characteristics of a case study the interviews go in-depth on the topic of algorithmic fairness and the context revolves around how companies already are approaching and implementing fairness in their ML systems, or how practitioners and regulators would like to see fairness implemented and managed. Several headlines showcasing algorithmic unfairness in the real world also add support to the idea of the case as pre-existing [20, 72]. The interviews were conducted in a semi-structured manner, which sought to take advantage of the natural setting and facilitate for relaxed conversations. People were interviewed directly in order to achieve first-hand knowledge and insight about the case. Multiple relevant industry representatives were interviewed to obtain a wider and more diverse perspective. The information gathered from these interviews is regarded as qualitative data.

4.1 Research Questions

The research questions are designed to convey the objectives and goals of the study, this section elaborates further on the motivation behind the decided research questions. From the systematic mapping study in Chapter 3 it was observed that recent research on algorithmic fairness has had a significant increase, with numerous techniques and solutions proposed, such as frameworks and metrics. Classification algorithms in particular have received increased attention, as their binary output provides a straightforward way of measuring unfair outcomes. The real world poses many problems and decisions however, many who cannot be reduced to binary problems or easily measured by fairness. Meanwhile, discoveries of algorithmic unfairness and discrimination appear in fields such as welfare [15], healthcare [56], jurisdiction [3], and education [4].

The increased attention that algorithmic fairness has received, has resulted in a lot of discoveries as well as considerations in a subject that was somewhat unheard of before, and which machine learning practitioners now need to adress and take a stand on.

Together this led to an aspiration to explore the presumed gap between research proposals and implementation by real-world practitioners. The following research questions have been motivated by this:

> **Research Question (RQ)1: How do companies approach and implement algorithmic fairness?**
>
> - **RQ1.1: What are the main challenges that companies face when implementing algorithmic fairness?**
>
> **RQ2: How do companies identify potential sources of bias in their algorithms, and what strategies are used to mitigate these biases?**
>
> **RQ3: What are the factors that facilitate or hinder the implementation of algorithmic fairness?**

4.2 Data Collection Procedure

The data generation of choice was semi-structured interviews, as it is an appropriate method for qualitative data analysis [55]. This approach allows for the discovery of unforeseen information due to more room for expression compared to structured interviews. This type of interview also complimented the time constraint of the project and fitted the availability of interviewees, allowing interviews to be conducted whilst also recruiting new participants.

The author spoke to all subjects over video calls, and the data collection process can be considered a first-degree data collection technique. This technique requires effort but is beneficial because it allowed the author to control how and what data is collected. More specifically to ensure that interview questions are answered and what follow-up questions to explore, in order to discover new directions. Qualitative data is often rich and broad, as opposed to precise, and it was, therefore, important to get the right interpretation of the responses by the interviewee, which can be challenging when analyzing qualitative data.

All interviews followed the interview guide (Appendix C), which was structured into general questions first, such as background, role, and fairness impressions, and then asking about their approach, where questions were asked differently depending on how systematic their approach was. Participants signed consent forms (Appendix B). All interviews were performed between February and April 2023.

Table 4.1 presents the details for each interview. All of the interviews were conducted through Microsoft Teams[1], and transcribed shortly after completion. In order to improve the quality of the next interview in regards to the questions asked, and to emphasize the most important topic, initial impressions and reflections were written down after each interview.

IDs	Company	Role	Size	Duration
R1	State-owned enterprise	Data Scientist	20000	55 min
R2	State-owned enterprise	Lawyer	20000	50 min
R3	State-owned enterprise	Data Scientist	500	50 min
R4	Private Research	Research Director	100	50 min
R5	Insurance	Director	4000	50 min
R6	Private Corporation	Data Scientist	500	45 min
R7	State-funded enterprise	Senior Advisor	100	55 min
R8	Private Corporation Company	Lawyer	100	55 min
R9	State-owned enterprise	Technologist	100	40 min

Table 4.1: Case Interviews

4.2.1 Recruitment of Relevant Participants

Relevant participants are people who work in companies and organizations where algorithmic fairness is implemented or of concern. People who met the following criteria were relevant for inclusion in the study:

- Develops, maintains, consults, or audits algorithmic decision-making systems
- Implements or considers algorithmic fairness in algorithmic decision-making systems

Participants were primarily selected in three ways. One way was by contacting those who had participated in public conferences where algorithmic fairness was a topic, or similarly had published articles or academic papers where algorithmic fairness was a topic or subtopic. The second way was using the author's existing network. The author didn't have any concrete ties to people or companies specifically working with algorithmic fairness, but instead had contacts who did. By utilizing their network, the author was able to obtain relevant contacts and potential pointers

[1] https://teams.microsoft.com/

this way. The third way was using LinkedIn[2] to search for topics like 'algorithmic fairness' and similar, to find people who worked with machine learning and AI in companies where it would be logical for fairness to be a part of their projects. Public sector companies, insurance companies, banks, healthcare companies, IT-consultancy companies as well as law firms with AI specialists are examples of companies where potential interviewees could be found. Researchers who participated in industry projects were also considered relevant.

There are numerous people working with algorithmic decision-making in Norway, but not necessarily with a focus on algorithmic fairness. Algorithmic fairness is often considered as one of several aspects within the responsible AI context [39]. Finding relevant participants was not as straightforward as hoped, and the objective was to find 6-12 people who worked on projects where algorithmic fairness in algorithmic decision-making systems was a topic. This could be people with varying backgrounds and roles, and ensured a broad unit of analysis, from a range of industries. This was deemed fitting as the SMS revealed that research on the topic is done both on domain-specific and generic research areas. In Section 4.3, a textual description of each interviewee is given.

4.2.2 Personal Data

To protect participants and avoid unnecessary worries, it was desirable to minimize the transfer of sensitive information and personal data from interviewees in particular. In order to communicate and set up meetings it was however necessary to obtain the name and email of each participant. Interviews were conducted online using Microsoft Teams, using the built-in functionality provided in order to record voice and audio. Teams also offer automatic transcriptions of meetings, this transcription is saved as a text file.

Some additional information was also collected from the participants, this included:

- **Years of Experience**
- **Working Title:** Current title of the participant.
- **Project:** Non-revealing information about projects the participant has been involved in.
- **Size of Company:** The relative size of the company.

This information is used to conduct the research and to produce the thesis. Because of this, necessary measures are taken in order to assure that no participants are identifiable from the data in the research.

Once the project is completed, all personal data will be deleted and personal data will not be accessible through the publication of this thesis. Participation in the project is entirely voluntary, and participants have every right to withdraw at any time and can withdraw their consent form without any further notice. This information was also highlighted in the Information Letter. The Information Letter also contains information about the purpose of the research, as well as the plan for how their personal data will be dealt with. Prior to the interviews, the author also told participants that they simply had to reach out if anything was unclear or if they had any questions. The letter is based on the template provided by Sikt, and the full version of the letter can be found in Appendix B.

[2] https://linkedin.com/

4.3 Case Descriptions

This section presents the participants in the case study, including their title and experience, as well as information about the company they work for such as the size, area, and a description of the relevant systems or projects they have partaken in. Descriptions are made as accurate as possible without exposing sensitive information about the participant, nor the company they work for. Case descriptions are provided so that transferable results are possible and to better understand the results. It's also emphasized that to the author's knowledge, in general, there are very few that work on *only* algorithmic fairness, as potentially a researcher could. Instead, practitioners are usually developers or maintainers of AI systems and solutions, where algorithmic fairness is an important aspect. It's also added that the AI systems described are not only algorithmic decision-making systems but also automatic decision-support systems.

R1 Data Scientist R1 works in a company with around 20000 employees. R1 works as a Data Scientist and has done so for the company in the last 5 years. The company R1 works for is a public agency that partakes in several fields, the most relevant to R1 is welfare. The company is working on an AI project that focuses on predicting the progression of individuals on sick leave. The system's prediction would not be the final output but instead, be given to a case manager who would use this information along with other information in order to make a final decision. R1 is thus concerned with algorithmic fairness in regard to an automatic decision-support system that would affect people's life. R1 also works on developing other AI systems, but this is the one that is the most relevant. R1 also follows the literature that is done on algorithmic fairness, such as by researching different toolkits that are available.

R2 Senior Advisor. R2 works in the same company as R1 and is a lawyer. They work with the same projects as R1 does but have a different role, as R2's main role is to give legal advice to different teams using machine learning. This includes making sure that the machine learning systems follow the law, and requirements such as fairness, explainability, transparency, and privacy. Assuring that the translation between law and code is correct is one task that is particularly important. R2 expertise does not lie in the technical aspects of algorithmic decision-making, instead, they use their legal expertise in order to oversee the translation between law and code that is done by developers and data scientists.

R3 Data Scientist. R3 is educated as a sociologist but now works as a data scientist in a company with around 500 employees. R3 works in a company that specializes in auditing and controlling various systems and solutions. They work in the company's artificial intelligence department, where tasks include auditing machine learning systems and algorithms, and this is where the relevance of algorithmic fairness comes from in the work that R3 does. R3 follows the literature regarding algorithmic fairness and other publications about artificial intelligence and has also authored papers about artificial intelligence and fairness. They work both on implementing machine learning in their own systems and processes and also auditing other companies' use of machine learning and algorithms. Certain projects R3 has worked on were in relation to analyzing and auditing machine learning algorithms and checking for certain biases.

R4 Research Director. R4 works as a researcher specializing in machine learning in a company with around 100 employees, R4 has 20+ years of experience. R4 works tightly with both companies and research institutions. R4 stays updated with algorithmic fairness research, and the increase in literature is part of why R4 has taken a special interest in the field. R4 often works on projects where R4 or R4's team only has partial responsibility such as only being in charge of the technical implementations, whereas another team has the superior responsibility, which may include deciding the fairness definition. Their task in these projects is usually to design the algorithm used in a solution and implement fairness accordingly.

R5 Department Director. R5 works for an insurance company with around 4000 employees and has studied economics. They work as a department director and has 10 years of experience. In order to process insurance claims and decide insurance premiums, the company employs thousands of machine learning models. R5 has a long experience with insurance and the use of machine learning within the insurance context. Algorithmic fairness is vital for R5 along with other aspects of RAI. Fairness is a relatively new concept in regards to the use of algorithms, but at the same time seen as very important, and a key factor for the future in terms of reputation and business value.

R6 Data Scientist. R6 has worked with algorithmic fairness both as a researcher as well as working as a Data Scientist. They work for a company that makes safety software and has around 500 employees. The current company of R6 is in the process of implementing more and more machine learning in order to streamline their solutions, although it's still at an early stage. R6 has previous experience working for an IT consulting company, where among other things they would provide solutions for implementing algorithmic fairness in AI systems. R6 also follows the literature and has attended several conferences on fairness in AI. Through this work as well as staying up to date with the literature, R6 has a good overview of existing solutions and toolkits.

R7 Senior Advisor. R7 has a background in the social sciences and is now working as a senior advisor in a company with around 100 employees. They have more than 5 years of experience working with the use and effects of AI. R7 works for a company specializing in consumer rights, such as ensuring fair treatment when a system uses algorithmic decision-making. R7 thus provides a different view on algorithmic fairness, as they "represent" those affected by algorithms, as opposed to those who design and deploy them. As a consequence of this, R7 doesn't always have all of the tools for detecting algorithmic unfairness at their disposal, as they may not have all the data or outcomes available. Instead, they employ different methods for bias detection, such as algorithmic auditing and unsystematic approaches.

R8 Lawyer. R8 is a lawyer who specializes in AI. R8 has worked at their current company for 3 years and the company has around 100 employees. R8 follows the research that is done and has a particular interest in algorithmic fairness. They work with client companies that wish to ensure that their AI systems are in line with legal regulations, which include ensuring algorithmic fairness. R8 is concerned with how the use of artificial intelligence challenges legal principles, and how bias in algorithms is a challenge to the principle of justice.

R9 Data Scientist. R9 works as a Data Scientist for a company with around 100 employees specializing in digitalization and privacy. R9 has 10+ years of experience working with AI for different companies. Among other focus areas, the company that R9 works for leads artificial intelligence projects where different companies can try out and evaluate their systems. These projects often revolve around privacy and RAI, and around half of the projects are also concerned with fairness. R9 has partaken in these projects where algorithmic fairness is important, and the projects operate in several different contexts such as healthcare, welfare, and surveillance, where both technical and organizational solutions have been proposed to mitigate bias and implement algorithmic fairness.

4.4 The interview Process

Before an interview took place, a participant would receive information about the project and the interview through an 'Information Letter' (Appendix B) that was sent via email. In addition, this letter contained detailed information about how data about a participant would be stored and processed. At the end of the letter was a consent form, each participant had to sign this in order to partake in the research. Once a suitable date was found, an interview was scheduled and an invitation to a Microsoft Teams meeting was sent.

When participants were recruited, they were given some instructions about what to expect the interview to be about. In this way, they would have some time to think about their views regarding the topic. Sending information about topics and questions to allow the interviewee to prepare can have a positive effect [55]. Another probable benefit of this is that it can alleviate some of the pressure an interviewee may feel. Algorithmic fairness is regarded as a new and emerging topic, meaning that in general, interviewees would not have a complete technical, societal, and organizational competence about algorithmic fairness, nor was this required as the goal was to explore implementation and approach, rather than only measure the theoretical knowledge level of practitioners.

Due to time constraints and geographical reasons, each interview was conducted digitally using Microsoft Teams. Throughout the interview process, the author participated as the sole interviewer, responsible for all sections of questions. The interview simulated a typical conversation in an attempt to ensure that the interview felt less rigid and artificial for the interviewee. By being the sole interviewer, the author aimed to maintain consistency in the interview process and minimize potential biases that may arise after conducting multiple interviews. All interviews were conducted in Norwegian, as that was what the participants wanted, and allowed interviewees to be as comfortable as possible while facilitating in-depth questions.

Each interview started up with a greeting and an appreciation for the participation of each interviewee, followed by an introduction of the author, a bit of background information, such as why this study was being conducted and the context around it, as well as informing of the semi-structured nature of the interview. This was followed up by some small talk to get the conversation going and hopefully make the interviewee feel relaxed. Oates states that thrust is essential, and particularly of interest if one is going to ask more sensitive questions [55]. Albeit sensitive questions don't make up the majority of the questions asked, they do involve reflecting on moral principles and considerations, so ensuring the well-being of a participant was of the essence.

After introductions were in order, each interviewee was reminded about why their answers and insight were valuable to the study. Before recording started the participant was informed that they would now be recorded, both audio and video. Oates states that informing participants that they are being recorded is pivotal for online interviews [55]. Thereupon, the interviews would follow the interview guide. As participants were informed that there was nothing wrong with taking their time and giving thorough answers, some questions were skipped as answers would overlap. In addition, sometimes the order of questions would change, if that could benefit the flow of conversation. It was also sometimes beneficial to ask follow-up questions that were not in the guide as issues emerged from the participant's answers. These are characteristics of a semi-structured interview, and have an emphasis on discovering rather than checking [55]. Apart from general questions at the beginning of each interview, the author asked open-ended questions as opposed to closed questions, as proposed in Oates' work [55]. The experience in general was that participants benefited from this, as it allowed them to freely express their opinions and share their experiences. The interviews concluded by asking the participants if there was anything they would like to revisit or add additional information to, as suggested by Oates [55]. At the conclusion of the interview, the author made an emphasis to once more thanking participants for taking part in the study [55].

4.5 Ethics

Ethical considerations must be considered when planning and conducting empirical research [55]. NTNU has appointed Sikt (Norwegian Agency for Shared Services in Education and Research) as their Data Protection Official for Research. All personal data collected for this thesis has been declared to Sikt. A consent form was created based on the template provided by the Norwegian Agency for Shared Services in Education and Research (Sikt)[3]. This was done in order to ensure that the research was in line with Norwegian law and the policies for the protection and management of both intellectual property rights and physical material at NTNU. This serves as quality assurance for the collection of personal data related to this thesis. Approval for the application to Sikt and the consent form is shown in Appendix B. The application was approved prior to starting the interview process. All who participated in the interviews were informed that participation in the study was entirely voluntary and that they had every right to withdraw from the study at any moment without further notice.

[3] https://sikt.no/en/home

Chapter 5

Results

This chapter presents the findings from the case study as themes relevant to answering the research questions. The chapter proceeds as follows: Section 5.1 presents how the thematic analysis was applied. Section 5.2 presents the themes that were generated paired with relevant quotes from participants, along with a higher-order grouping of themes, placing them as either technical or social. Finally, section 5.3 discusses the reliability and validity of the results.

5.1 Analysis Procedure

As stated, the interview data is regarded as qualitative, as it consists of text rather than numeric data. The analysis procedure will thus be a qualitative data analysis on the data that was gathered through a case study, stemming from recordings of the interviews. Thematic analysis, a method for identifying, analyzing, and documenting recurring patterns found in data, is applied to qualitative data [10]. It provides a basic level of organization and offers an in-depth description of the dataset. The thematic synthesis process aims to answer the research questions. There are several strategies and guidelines for carrying out qualitative data analysis, Kiger et al. describe a six-step framework [41], whilst a 5 step thematic synthesis approach consisting of data extraction, data coding, code to theme translation, model of higher-order themes and system trustworthiness assessment is suggested by Cruzes et al.[18]. Oates states that there are no strict rules [55], but his advice and guides on approaching qualitative data analysis will be followed to a large extent.

5.1.1 Transcription

Oates's first step in qualitative data analysis is data preparation, which consists of preparing the data for analysis by getting it stored in the same place in the same format[55]. All of the interviews were conducted through Microsoft Teams, and recorded through Teams's recording feature. Teams also offer automatic transcription which gets stored as a text file. After each interview that data outcome was one video file containing the recording of the interview, and a text file of the transcription. Each interview was transcribed using both automatic transcription tools and manual transcription. The manual transcription was necessary because automatic tools aren't as good for the Norwegian language as compared to English, especially when there are factors such as dialects, abbreviations, or the occasional use of English words. By having all the interview data in the same format the information becomes easier to access and read through. Braun et al. refer to this step as "familiarizing yourself with your data" [10], and that the transcription of verbal data allows for familiarization. The step can often be seen as tedious, but it serves as the bedrock for the analysis [10]. As interviews were conducted in Norwegian, they were also transcribed and analyzed in Norwegian. As it's challenging to guarantee that no one is recognizable from an entire interview lasting up to an hour, complete interviews are not included in the thesis. Certain pieces of the interviews, such as important quotes, were translated into English. These quotes are used to support results and findings.

5.1.2 Initial Reading

Interviews were transcribed shortly after completion to capture the actual meaning of the answers given by the interviewees. The first step of the analysis was to read through the text, which consists of all the conducted interviews. The goal of this reading is to generate introductory ideas and pinpoint potential patterns in the collected data. This step can also be seen as an exploration of the data, in order to establish a general sense of the information one is dealing with, before it is broken down into parts [17].

5.1.3 Coding Process

In order to further process the text, a process called coding is performed, it consists of segmenting and labeling text. The object of the coding process is to make sense of the text, divide it into segments and label them with codes, these codes are then examined regarding overlap and potential redundancy [17]. The process also involves selecting which data to use, and what to disregard. The disregarded data is thus not used to directly provide evidence for later themes.

The author applied an iterative coding process as it suited the time constraints by allowing for both data collection and data analysis at the same time [17]. Data from the first four interviews were coded in the first iteration. In the second iteration, the four next interviews were coded. The final interview was coded in a third iteration as that interview happened sometime later than the other interviews. In the second and third iterations, the data from the interviews were classified into existing codes or generated new codes, and in the end, resulted in 42 codes.

5.1.4 Translate Codes into Themes

By combining several codes, one can create themes. These themes are generally broader than the identified codes. The goal is to find themes that give helpful information about the data regarding the purpose of the analysis. A code that is insightful enough could therefore also become a theme by incorporating other similar codes, and some codes are discarded if they are deemed irrelevant or too vague. In order to ensure the correct context each case was analyzed separately first, and then combined with similar codes from other cases. Themes form a vital element in the qualitative analysis as they form a major idea of the collected data [17].

After the first generation of themes were created it is important to review them [10]. By comparing the themes to the data one can ensure that the themes are useful and give accurate representations of the data. Problematic themes can be split up, combined, discarded, or made into new ones to make themes more accurate and useful. One may also move a code from one theme to another. Finally, each theme is given a precise name that formulates what the theme consists of and what it represents in the data gathered from the interviews. In order to not introduce overlapping terms or contribute to more confusion, theme names are adapted from the literature when possible.

5.2 Presentation of themes

This section presents the themes and how they came to be using quotes from the interviews. Themes are categorized in a higher order as either Technical or Social factors regarding their main contribution, to indicate where they are most relevant in regard to the sociotechnical view of algorithmic fairness. In the end, a total of 14 themes were generated. The final composition of themes is shown in Figure 5.1. A definition of each theme is presented in Table 5.1.

5.2.1 Technical

Formalism Trap. Depending on the company's obligations and its goal, the fairness definition in a machine learning system will typically stem from either the law or from the perspective of stakeholders, possibly both. Either way, companies still have to be able to map between this written definition and how it looks in code, but translating between what one thinks is fair and implementing it in a solution is not always straightforward. R6 shared the struggle to properly capture the nuances of fairness in a technical solution.

> R6 - "*Things make sense when you're discussing them, but when you're supposed to write it in code, you understand that you can't always write this as an algorithm. It's very difficult.*"

R1 also shared their experiences with formulating fairness:

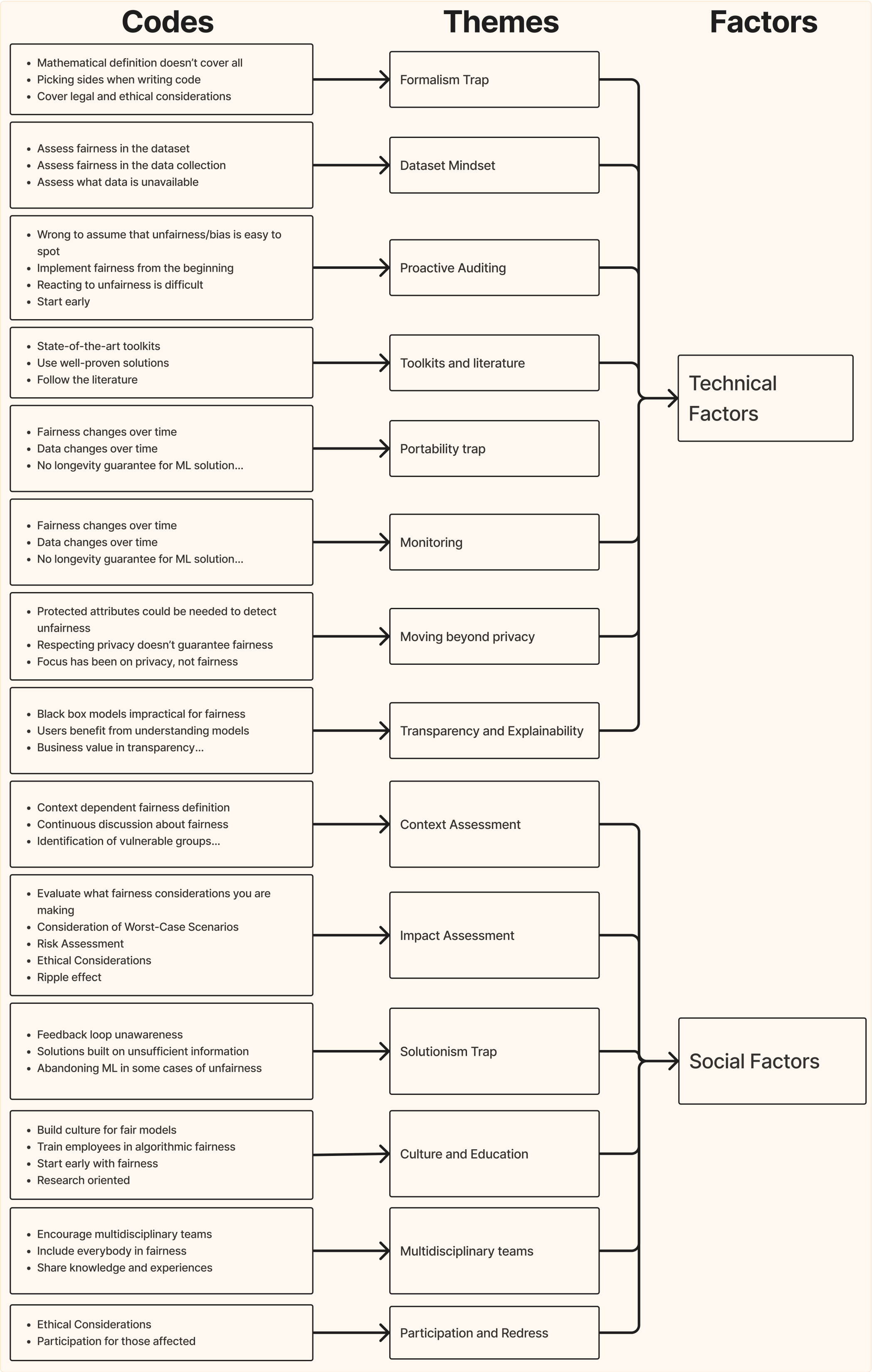
> R1 - "*An unsolved problem for us is to map between mathematical fairness metrics and the legal understanding in the given context.*"

The failure of accounting for the full meaning of social concepts is referred to as the *formalism trap* and is common when trying to model a social problem [66].

If the system isn't necessarily bound by legal obligations, companies still have to be able to communicate their mathematical definitions of fairness.

> R5 - "*You need to be able to translate the mathematical fairness definitions, into a way that a domain expert can understand and see what moral principles apply.*"

If one can't do this then one will have little credibility when claiming your system is fair.



Codes
Themes
Factors
Mathematical definition doesn't cover all
Picking sides when writing code
Cover legal and ethical considerations
Formalism Trap
Assess fairness in the dataset
Assess fairness in the data collection
Assess what data is unavailable
Dataset Mindset
Wrong to assume that unfairness/bias is easy to spot
Implement fairness from the beginning
Reacting to unfairness is difficult
Start early
Proactive Auditing
State-of-the-art toolkits
Use well-proven solutions
Follow the literature
Toolkits and literature
Fairness changes over time
Data changes over time
No longevity guarantee for ML solution...
Portability trap
Fairness changes over time
Data changes over time
No longevity guarantee for ML solution...
Monitoring
Protected attributes could be needed to detect unfairness
Respecting privacy doesn't guarantee fairness
Focus has been on privacy, not fairness
Moving beyond privacy
Black box models impractical for fairness
Users benefit from understanding models
Business value in transparency...
Transparency and Explainability
Technical Factors
Context dependent fairness definition
Continuous discussion about fairness
Identification of vulnerable groups...
Context Assessment
Evaluate what fairness considerations you are making
Consideration of Worst-Case Scenarios
Risk Assessment
Ethical Considerations
Ripple effect
Impact Assessment
Feedback loop unawareness
Solutions built on unsufficient information
Abandoning ML in some cases of unfairness
Solutionism Trap
Build culture for fair models
Train employees in algorithmic fairness
Start early with fairness
Research oriented
Culture and Education
Encourage multidisciplinary teams
Include everybody in fairness
Share knowledge and experiences
Multidisciplinary teams
Ethical Considerations
Participation for those affected
Participation and Redress
Social Factors

Figure 5.1: Codes and Themes generated through thematic analysis

Dataset Mindset. Most research conducted on algorithmic fairness views the dataset as fixed and instead focuses on algorithmic methods that can mitigate bias [13]. However, in many cases, ensuring that the data is collected with fairness in mind could both be easier and more beneficial. Companies expressed the importance of being aware of potential biases in the data they collect and understanding how data is collected, such as who collects it and biases that might lie in the data collection.

> R7 - "*There is always a risk of biased data collection.*"

This often involves working closely with domain experts. An example is to be aware of what your data lacks, for instance, underrepresented groups in the dataset. This resonates with other studies, Holstein et al. found that most teams look to the training data as opposed to the machine learning algorithm when seeking to improve fairness [37].

The quality of data is a key aspect in all of machine learning. Errors, outliers, and noise all make it difficult for a model to detect underlying patterns to perform well [31]. The better data one has the better a model can predict, the "garbage in, garbage out" is a rather famous saying among computer scientists. Machine learning models can't ask for more or better information, instead, flaws in the data are accepted, and even the newest algorithms become helpless when data quality is poor. And so to prevent algorithms from forwarding biased outcomes companies need to ensure that their data quality is high.

Heavily linked to data quality is to have enough data. Collecting the right data is also an important aspect. When discussing lessons learned, interviewees shared examples where the model's prediction seemed right based on the data they had, and it was not until they sought out a more diverse demographic that they realized the prediction was wrong and discriminatory, and that the data basis was insufficient.

> R9 - "*Understanding that we may actually need to use sensitive attributes to achieve fairness was a bit of a learning curve for us.*"

Proactive auditing. One aspect of working with algorithmic fairness that participants shared is that you can't simply begin considering it when your model is already deployed and affecting people. Starting early with fairness is essential for success.

> R1 - "*When working with fairness you need to start early, not just because of legal considerations, but also because it affects the design and product development of the solution.*"

Starting early also has the benefit of reducing the need for costly and time-consuming revisions further down the line of product development. Retroactively integrating fairness considerations in an AI system that is operational can be complex and costly, and it's generally more cost-effective to prevent unfairness in the first place rather than to face the aftermath of algorithmic bias. Multiple examples of negative headlines and compensation claims that damage reputation serves as motivation to start early.

Spotting bias in the data or in the outcomes isn't always as straightforward.

> R9 - "*We have seen that methods for detecting bias are often very qualitative, and based on discretion. ... its wrong to assume that it's easy to spot discrimination and bias. A lot of biases are you don't see, and there is a need for systematic methods, and quantitative, data-based assessments. One of our conclusions so far is that these methods aren't very widespread yet.*"

Bias through proxies, or unfair treatment only observable through intersections make it difficult to be sure that the system is actually fair, and poses a challenge for practitioners.

These aspects fall under the theme called *proactive auditing*, which emphasizes the need to make fairness considerations from the beginning, as opposed to later having to mitigate unfairness and having mechanisms in place for analyzing the system.

Toolkits and literature Most participants noted that they stayed updated on algorithmic fairness research and that their fairness entry point often was from the literature.

Several toolkits and frameworks have been created to help practitioners make the technical implementation easier and more robust. R1 for instance shared how they used the Aequitas fairness three (Section 2.5.1) when choosing a metric to evaluate their system:

> R1 - "*Choosing a metric isn't a trivial task but there are certain guidelines you can follow, we primarily used the fairness tree developed by Aequitas*"

R6 elaborated on how they picked their toolkit of choice:

> R6 - "*We did a little survey on available toolkits, and ended up using AI Fairness 360 because it supported most algorithms and had useful documentation and references*"

Portability trap Being able to apply a solution for different problems is usually sought after as it can save on development costs. However, a solution that was created for a specific social context may work differently when applied in another context.

Holstein et al.found that since fairness can be context and application dependant there is a need for domain-specific processes and tools [37]. R3 shared how they developed their own domain-specific framework in order to audit machine learning algorithms.

> R3 - "*We made our own framework for auditing machine learning models based on our own experiences ... where fairness is covered, although in general terms.*"

creating solutions that are aware of the context it's applied in is therefore crucial. The failure to understand how different social contexts may need different algorithmic designs is called the *portability trap* [66].

Monitoring The need for mechanisms that can monitor fairness in advanced systems was expressed by several participants. Fairness is not a static concept and may change over time. Similarly, the training data that goes into a model may change over time, altering how a model makes predictions. R9 expressed the importance of monitoring solutions, as opposed to leaving the model alone after it's been deployed.

> R9 - "*There is no guarantee that a (machine learning) system that performs well today will perform well in 6 months.*"

As well as it being difficult to guarantee the longevity of a machine learning solution, fairness being a social construct means it is susceptible to change, R6 pointed out the need to be aware of this:

> R6 - "*Fairness can change over time, so you need a process that monitors the solution so that it's correct over time*"

Moving beyond privacy. Practitioners usually don't have fairness as the only requirement for their AI system, instead, it is one of several RAI elements. One element that is also often important, is privacy. However, ensuring privacy is not the equivalent of ensuring fairness. R9 shared thoughts on how they focus on data minimization, which means limiting the information the model needs, while still being able to perform fair and accurate.

> R9 - "*We have a focus on minimizing the model,.. not using more sensitive information than is required.*"

The assumption that if the AI is sheltered from sensitive attributes it cannot be biased is called fairness through unawareness. This unawareness trap is important for companies to be aware of, especially when privacy is also of concern. In fact, having access to sensitive attributes may have a positive effect for allowing the model to achieve algorithmic fairness, but reduce privacy, leading to a privacy-fairness trade-off [12].

Transparency and Explainability. Another challenge is the 'black box' nature of many machine learning models. These models, particularly deep learning models, often lack transparency, making it difficult to understand how they make decisions. This opacity hinders efforts to identify and correct bias within these algorithms. The desire for precision and speed of systems leads to the use of more machine learning, especially 'black box' models. Providing sufficient transparency while maintaining the effectiveness of complex models remains a significant challenge.

> R5 - "*The need for more speed and accuracy generates the need for more usage of machine learning models, especially black box models.*"

5.2.2 Social

Context Assessment In order to implement fairness you need to have a definition of fairness. When considering what is fair, companies have several considerations to make. This challenge stems from the discordant nature of what fairness means. As discussed it depends on many factors and can change over time. So choosing a definition can depend on politics or philosophical conviction, but also on the type of result one wants to achieve. This particular part raises another question, *fair for who?* fair for stakeholders? fair for all users? fair for certain users? This is a difficult challenge and demanding exercise for practitioners. R2 works on ensuring that machine learning systems meet legal requirements, and have thus used the law to translate between social and mathematical fairness.

> R2 - "*My starting point will always be the law. ... and depending on the context, we have to see what the relevant laws say.*"

R5 informed on how it's difficult to conclude and agree on a fairness definition:

> R5 - "*You can argue both ways whether something is fair or not.*"

Impact Assessment Once a fairness definition has been decided, it is equally important to understand the impact of this fairness definition. This is a difficult task as properly understanding the consequences requires one to consult domain experts. R6 pointed out that this communication requires a thorough understanding of fairness definitions. This understanding must be translated into terms comprehensible to the domain expert. This involves crafting compelling arguments, providing examples that describe potential outcomes of different actions, and showing the consequences of inaction. The process invariably triggers extensive debates, as individuals come forth with differing perspectives. The risk of different scenarios needs to be mapped out and agreed upon. R6, a data scientist stated that this issue is often more challenging than creating the technical solution.

> R6 - "*Talking to people, that is what is difficult. Always, all the time. when it comes to changing processes, and convincing people, that's where the complexity lies, I think. ... Don't just say, should I choose this framework, or that framework, or that library, to implement the algorithm. That's the easiest part.*"

Understanding who the vulnerable groups are in a project can also be difficult, and discrimination can be hidden when looking at features in isolation.

> R5 - "*Discrimination at the intersection is a worry to us because we have no guaranteed way of preventing it.*"

Solutionalism trap Interviewees were asked what type of machine learning systems they work on / have worked with, yet some expressed the need to also be open to the possibility that some problems aren't ready for AI solutions. R1 stated that if their team wasn't able to train a model that is both accurate and fair, then moving away from AI could be considered.

> R1 - "*Sometimes the best solution may be to not use AI for this problem.*"

In projects where different teams work more or less isolated on different parts of the system, such as when several companies are involved, it can be difficult to see the bigger picture and understand if AI is suitable for the project. R4 shared that their team sometimes only focuses on the technical aspect:

> R4 - "*Our team may be tasked with designing the algorithm, but not involved in discussing the fairness definition.*"

Ignoring that the optimal solution doesn't always involve a technical solution is referred to as the *solutionalism trap.* [66].

Culture and education Algorithmic fairness has received increased attention within the research community in the last years [22]. As AI becomes more and more widely used, interviewees proclaimed the need for more education about both understanding AI in general, but especially about algorithmic fairness.

> R8 - "*The lack of knowledge about algorithmic fairness is a worry. ... building a culture for fairness is important.*"

Participants felt that there is a need for building a culture around algorithmic fairness and fair models and that it becomes natural to make considerations regarding algorithmic fairness.

> R5 - "*Regulations like the EU AI Act help push us to have procedures in place for ensuring algorithmic fairness*"

> R7 - "*Hopefully we can get to a point where fairness is something you keep in mind and address, similarly to GDPR.*"

Multidisciplinary teams. The value of multidisciplinary teams in the context of algorithmic fairness is substantial. AI and ML applications exist at the intersection of technology, ethics, law, social sciences, and business strategy, hence, demand a holistic approach that transcends disciplinary boundaries. Engineers and data scientists provide the technical expertise required to develop and refine AI systems. However, ethicists, sociologists, psychologists, and legal experts can provide valuable insights into the wider implications of these technologies, contributing to a more comprehensive understanding of fairness. They can help identify potential social, ethical, and legal pitfalls, propose alternative perspectives, and suggest measures to mitigate biases. Therefore, the presence of a multidisciplinary team can foster a more nuanced understanding of algorithmic fairness, leading to AI and ML systems that are not only technologically advanced but also ethically sound, legally compliant, socially responsible, and strategically aligned. Approaching fairness through multidisciplinary teams provides value.

> R1 - "*When you're going to attack this, it's smart to have a multidisciplinary team so that you can look at it with different eyes, not just a data scientist, or just a lawyer, but that there is a bit of diversity in these teams, and we think that's good, in the long run. I think that gives you more perspectives, and you get to raise the issues.*"

Participation and Redress Involving stakeholders in the design, implementation, and evaluation of algorithms ensures their perspectives and concerns are heard and considered. Actively engaging those who are affected can ensure that their perspectives and concerns are heard and considered.

> R5 - "*We've actually engaged expert consultants with hands-on experience that can help us understand what fairness considerations we should be making.*"

However, this might not be suitable or doable in all cases such as privacy limitations or in cases where one doesn't have the necessary means or resources to interact with participants.

Factor	**Description**
Formalism Trap	Mathematical definitions eliminate the nuances of fairness.
Dataset Mindset	Both literature and to some degree companies focus on making algorithms fair, ensuring that the data is fair can be easier and more beneficial.
Proactive Auditing	Aspire to implement fairness from the beginning, instead of mitigating unfairness later.
Toolkits and Literature	Using state-of-the-art toolkits for technical evaluation and implementation and staying updated on research.
Monitoring	Maintaining that the outcomes are fair, and prevent bias and unintended consequences after initial development and deployment.
Moving beyond privacy	Understand that an AI system could respect privacy (by properly handling personal data) or be sheltered from sensitive attributes, but still be unfair (if it produces biased outcomes).
Transparency and Explainability	Fair algorithms should not be entirely black box. It's important for users to understand how the algorithm makes decisions, which requires transparency in its design and functioning. Explainable AI techniques can help achieve this.
Portability trap	Recognizing that reusing algorithmic solutions, originally designed for a specific social context, could lead to misinterpretations, inaccuracies, or potentially cause harm when implemented in a different context.
Context Assessment	Assessing the context of a system and how this affects how fairness is approached and defined, and who should be involved.
Impact Assessment	Assessing the impacts of an algorithm and potential negative outcomes necessitates understanding its social context and the varied notions of fairness within that system.
Solutionalism trap	Overlooking the possibility that the optimal solution may not always involve technology can lead to missteps.
Multidisciplinary teams.	Contribute to a comprehensive understanding of biases, ethics, and social implications in algorithmic systems. Foster critical thinking, challenge assumptions, and promote creative problem-solving, leading to robust and equitable solutions.
Culture and Education	Develop a culture for fairness. Necessary for developing domain-specific guides, algorithms, metrics, ethical frameworks, and case studies.
Participation and Redress	Affected individuals and communities should have the opportunity to participate in decision-making about algorithmic systems, and there should be mechanisms for redress if the algorithm causes harm.

Table 5.1: Descriptions of Themes

5.3 Quality Assurance

Performing Quality Assurance is important in order to ensure the reliability and validity of the thematic analysis. Since the interviews were performed and transcribed in Norwegian, sentences or sections had to be translated into English. Certain considerations as to be made in order to provide quality assurance, such as being aware of the meaning of certain terms and phrases, and then translating and using them in the correct context [28]. A particular case of this consideration is that the Norwegian language and its many dialects have many specific terms and phrases that may not have an equivalent English translation that carries the exact same meaning. Taking extra steps to identify these terms and phrases and understand their meaning was, therefore, necessary in order to achieve an accurate translation. Another quality assurance step incorporated into the translation process of this study was proofreading. In this context, proofreading entailed a comparative analysis of the Norwegian and English translations, sentence by sentence or paragraph by paragraph, to ensure that the intended meaning of the Norwegian text was fully conveyed in the English translation. An approach to ensure the coding quality is to test the coding reliability. One possible way of doing this is to have several researchers code the data and see if the results are similar. Since this is a one-person thesis it was not possible to implement, and is thus a limitation to this research as there is a risk for errors in the coding process [75].

Chapter 6

Discussion

This chapter is a discussion about the findings presented in Section 5. The chapter proceeds as follows: Section 6.1 answers research question 1 and its related sub-questions. Section 6.2 answers research question 2 and Section 6.3 answers research question 3. Section 6.4 presents *The Extended Sociotechnical Framework for Algorithmic Fairness*, based on relevant factors found through 9 interviews. Table 6.1 presents a summary of the main answers to each research question.

6.1 RQ1 How do companies implement/approach algorithmic fairness?

In order to implement algorithmic fairness, there are multiple considerations that companies must take. Each case will have a different context, data, effects on people, and end goal, but since they all share a motivation for fair algorithms there are some common denominators.

One aspect of working with algorithmic fairness is that you can't simply begin considering it when your model is already deployed and affecting people. Starting early with fairness is essential for success.

> R1 - "*When working with fairness you need to start early, not just because of legal considerations, but also because it affects the design and product development of the solution.*"

Starting early also has the benefit of reducing the need for costly and time-consuming revisions further down the line of product development. Retroactively integrating fairness considerations in an AI system that is operational can be complex and costly, and it's generally more cost-effective to prevent unfairness in the first place rather than to face the aftermath of algorithmic bias. Multiple examples of negative headlines and compensation claims that damage reputation serves as motivation to start early.

Evaluating the fairness of a machine learning model requires a definition of fairness, and thus companies must acknowledge that the definition they end up using may lead to other notions of fairness being impossible to achieve [44]. Thus, understanding the context and assessing the impact of the system is pivotal for companies' approach toward algorithmic fairness.

As mentioned, it is normally agreed upon that fairness is something one wants to achieve, and instead, the question is rather how it should be achieved. The interviews revealed that there are other limitations that interfere with achieving *true* fairness, such as legal requirements that need to be met. An example shared by interviewees is where the laws by policymakers interfere with achieving fairness.

> R2 - "*You can make a model and test its performance and fairness, but legislators can decide that certain groups should be prioritized over other groups, and then the model would have to be "unfair" first so that it complies with the law before it can be fair to other groups.*"

Changing laws and changing perceptions of what is fair show that fairness is a continuous topic. Solutions, therefore, need to be flexible and ready to adapt to changing requirements, while still ensuring equitable outcomes.

As stated, the research on algorithmic fairness has mostly been concerned with statistically defining fairness and then proposing methods and techniques to mitigate undesirable biases, in relation to these definitions [2]. Whilst practitioners to some degree also where concerned with providing statistically fair solutions, such as evaluating the false negative rate or setting a threshold for how much unfairness your definition of fairness can handle, the overall takeaway was that the social aspects was the main and most difficult part of algorithmic fairness. The reason for this was that what constitutes as fair is a concrete assessment depending on the context that requires a lot of knowledge to properly understand.

Some even stated that with today's toolkits, the technical aspect is a very small part of approaching or implementing algorithmic fairness.

> R6 - "*The technical implementation is a small part, the tools, and frameworks support you to check that your algorithm is implemented correctly and saves you from a lot of troubleshooting. It's a small part, but it's reassuring to have it in place.*"

How AI systems are allowed to operate seems to be changing, with proposals such as the EU AI Act putting a stop to intrusive systems that violate fundamental rights [69]. These types of regulations can make practitioners more aware of fairness and other aspects of RAI in their systems so that more companies become aware of the consequences of algorithmic unfairness. Going back to the social aspect being the main part of the implementation, practitioners also felt that it would be difficult to create legal requirements that guarantee that all outcomes are fair, instead, they saw it as more realistic that one is required to have made considerations for fairness, opposed to guaranteeing it.

6.1.1 RQ1.1 What are the main challenges that companies face when implementing algorithmic fairness?

There are several challenges that companies encounter when dealing with algorithmic fairness. Despite working on very different projects, the interviewees revealed that practitioners phase many of the same challenges. There is a good possibility that several other challenges exist, but that a good portion of problems relevant to current practitioners have been discovered.

Through the interviews, it was identified that improving the data quality and data collection is essential and can often be more realizable than improving the algorithm itself. Despite this, data quality is identified as a challenge because data may contain historical bias which the algorithm will reflect [63]. Similarly, data may be affected by the conscious or unconscious bias in the people who collect the data. All in all, there are multiple places where bias can seep into a system (see Figure 2.4), and that makes it challenging for practitioners to implement fair solutions.

Having enough data is also a challenge as unprivileged groups are often underrepresented. Pre- or in-processing techniques can help against this challenge by changing the sample distribution of sensitive attributes, or by balancing the different objectives of the ML model (i.e. fairness and accuracy) [11].

There are also cases where data isn't available, such as when all outcomes aren't observable. One version of this problem is typically referred to as counterfactuals, an example is getting a rejection for a loan, where one still doesn't know if the loan would have been paid back if it had been approved [71]. Expanding on this, if a loan was approved and later paid back, one does not know if the loan would have been paid back had circumstances been different (e.g., larger loan amount or longer loan term). This leads to a potential blind spot in the model, as there is no accurate way of measuring the false negatives. This is a machine learning challenge, that is intimately connected to algorithmic fairness, as models that are biased against certain groups could continue to reject candidates from that group. This further makes it difficult for a company to observe the counterfactual outcomes and assess whether the model is fair.

In an ideal world, no amount of unfairness and discrimination would be accepted, in the real world however, this is practically impossible. When using algorithmic decision making one can quantify the performance of the system, and it's unreasonable to expect 100 % accuracy, and likewise, a model may perform worse for certain groups than it does on average. Determining the threshold for how much unfairness your fairness definition accepts is therefore a difficult but necessary exercise. Once it's determined it's important to continuously monitor the system's up against that limit. Several participants noted how this is challenging, R8 stated the following:

> R8 - "*... the AI must be checked against this limit continuously. This is, for example, because the composition of the group of people the AI is used on can change, or the algorithm can become biased over time if it learns from and systematizes biases gradually*"

Sometimes, in order to create more fair outcomes, one needs access to sensitive attributes [12]. Despite this, other regulations, such as privacy rules can make it difficult to use these attributes in the training of a ML model. Because of this, companies may be more concerned with the privacy of their solutions, which can increase unfair outcomes, if one takes a fairness through unawareness approach. It, therefore, remains an open challenge for practitioners to balance the privacy and use of sensitive attributes and ensuring fair outcomes.

6.2 RQ2; How do companies identify potential sources of bias in their algorithms, and what strategies are used to mitigate these biases?

Bias can appear in several parts of a machine learning system, and thus there are several processes for identifying bias.

Bias may not always be so easy to spot, proxies can make it difficult to identify bias. Similarly, discrimination that only happens at intersectionality makes it difficult to understand when unfair treatment is happening.

For classification and regression problems one can use techniques such as feature importance to see what attributes the model utilizes the most in its prediction. Through these techniques, practitioners can learn potential biases in their model. It can also provide insight into their dataset as it shows what features are the most relevant, it is also possible for a domain expert to interpret the results and make suggestions on what data to gather more of or to gather different data. R3 was part of a project where they revealed bias by looking at the feature importance of the model.

> R3 - "*By using feature importance methods we were able to see the model being discriminatory towards gender, and pointed out that this unfairness should be looked into even though the project is in an early phase.*"

Experiences like this one serve as a reminder for practitioners that one should always look for these biases when AI is used in a sensitive environment.

Several participants pointed out that there is a lack of systematic methods for discovering bias and unfairness, and discrimination happening at the intersectionality of attributes is an example of bias that won't be discovered through only developer intuition and guesswork.

As mentioned, R7 offered a different perspective on this matter as they typically evaluate algorithmic decision-making systems externally. This leads to more unsystematic approaches, that have little guarantee of revealing unfairness. This raises an important challenge if one wants to advance toward fair algorithms. If algorithmic unfairness is difficult or impossible to discover by those who do not have access to all of the data, then evil-minded institutions would have very little incentive to ensure that their algorithms are fair, if there is no risk in doing so.

As machine learning models aren't always 100% interpretable, it is difficult to isolate where unfairness is happening or coming from. Relying on the developer's intuition for discovering unfairness is a risky strategy, but is often the chosen strategy, due to the lack of support to address the issue. A similar study by Holstein et. al also found that most industry practitioners rely on their intuitions, even though these were often found to be wrong [37].

6.3 RQ3 What are the factors that facilitate or hinder the implementation of algorithmic fairness?

Several factors can hinder the implementation of algorithmic unfairness. Rapidly evolving legal and regulatory landscapes surrounding algorithmic fairness can cause uncertainty for companies. When interpretations and guidelines are not yet well-defined or consistently applied, understanding and complying with laws and regulations related to fairness can be challenging. Furthermore achieving this understanding can be expensive if the company doesn't already have these resources. In general, a lack of resources can serve as a hindrance to algorithmic fairness.

Trade-offs related to performance can hinder fair outcomes. If machine learning is used by a company it is normally because it is more efficient and accurate than a human, and maintaining this performance can be essential to justify the resources invested in a model. Striking a balance between fairness and accuracy is therefore a challenging aspect that practitioners need to attend to.

An aspect that doesn't necessarily help practitioners advance toward algorithmic fairness, but can serve as a motivation in several contexts, is fairness as a selling point, with fairness adding business value. Fair algorithms enhance brand reputation and foster trust among customers and stakeholders. In an era where customers increasingly value ethical business practices, companies demonstrating a commitment to fairness can differentiate themselves in the market. Investing in algorithmic fairness is not just a matter of ethics and compliance, but also a sound business strategy that drives long-term value and competitiveness.

> R5 - "*We believe that implementing fairness, along with transparency and responsibility, will drive business value, and those who are best at it will have a competitive advantage. . . . fairness will become a selling proposition.*"

Fair algorithms can also lead to better and more inclusive decision-making. They can uncover and correct biases that may have traditionally limited business opportunities, such as in hiring, lending, or marketing. This leads to a more diverse and inclusive customer base and workforce, which are known to improve creativity, innovation, and profitability. Lastly, fairness can reduce the risk of costly litigation and penalties associated with unfair or discriminatory practices.

There have been numerous reports of algorithmic unfairness in the media [15]. Participants mentioned several different examples, and that these stories serve as a motivation for fair algorithms, as well as arguments for why one should be aware of the possibility of algorithmic unfairness.

Identified factors that are relevant for advancing toward algorithmic fairness are adapted into *The Extended Sociotechnical Framework for Algorithmic Fairness* presented in Section 6.4.

6.4 The Extended Sociotechnical Framework for Algorithmic Fairness

Based on the results from the thematic analysis, a framework for understanding how practitioners can advance toward algorithmic fairness has been created. The framework is based on the

RQ	Findings
RQ1	Implementing algorithmic fairness requires careful consideration. Starting early in the development process is crucial, as retrofitting fairness into an operational AI system can be complex and costly. Defining fairness and understanding the system's context and impact are essential. While achieving true fairness may have limitations due to legal requirements and varying perceptions, solutions should remain flexible to adapt to changing requirements while ensuring equitable outcomes. The social aspects of algorithmic fairness are considered the most challenging, emphasizing the need for a deep understanding of fairness within the given context. Both technical aspects and social considerations are important for practitioners. Regulatory proposals raise awareness about fairness and its consequences, prompting practitioners to be more mindful of fairness and other aspects of RAI. Legal requirements are more likely to necessitate considerations for fairness rather than guaranteeing it entirely.
RQ2	Bias can manifest in various parts of a machine learning system, making it important to employ processes for identifying bias. However, bias can be challenging to spot, particularly when it occurs through proxies or at the intersectionality of attributes, making it difficult to detect unfair treatment. Practitioners have several techniques to identify potential biases and gain insights into the dataset. The lack of systematic methods for discovering bias poses a challenge. Machine learning models' lack of interpretability complicates the detection of unfairness, relying on developers' intuition. Addressing these challenges is crucial to advance toward fair algorithms and ensuring the identification and mitigation of algorithmic unfairness.
RQ3	Several factors hinder or facilitate the implementation of algorithmic fairness. The rapidly evolving legal and regulatory landscapes can cause uncertainty and challenges in understanding and complying with fairness-related laws. Limited resources, including financial and technical capabilities, can also impede progress in algorithmic fairness. Balancing fairness with performance can be a challenge, as maintaining efficiency and accuracy is often crucial. However, there are compelling motivations for pursuing algorithmic fairness. Fairness can enhance brand reputation, build trust with customers, and differentiate companies in the market. It can also lead to better decision-making, foster diversity and inclusivity, and reduce the risk of penalties associated with unfair practices.

Table 6.1: Summary of answers to each research question

sociotechnical perspective introduced in Section 2.3. Expanding on the work of Sarker et al. the purpose of the framework is to improve the understanding of algorithmic fairness [65], forging a link between components and objectives. Figure 6.1 shows the proposed framework, and the Technical and Social Factors are explained in Table 5.1. The factors stem from the themes of 9 semi-structured interviews.

The framework is split into four main categories, consisting of General Techincal Factors, Case-specific Technical Factors, General Social Factors, and Case-specific Social Factors. This separation is done in order to understand what factors to some degree apply to every company concerned with algorithmic fairness, and what factors were found to be more case specific. The dividing of factors also makes it so that it's easier to use, and less overwhelming, whilst also emphasizing that there is no one-size-fits-all solution to algorithmic fairness [53]. The important aspect of context for this topic serves as a justification for this separation.

The idea behind the framework is to help bridge the connection between the technical and social

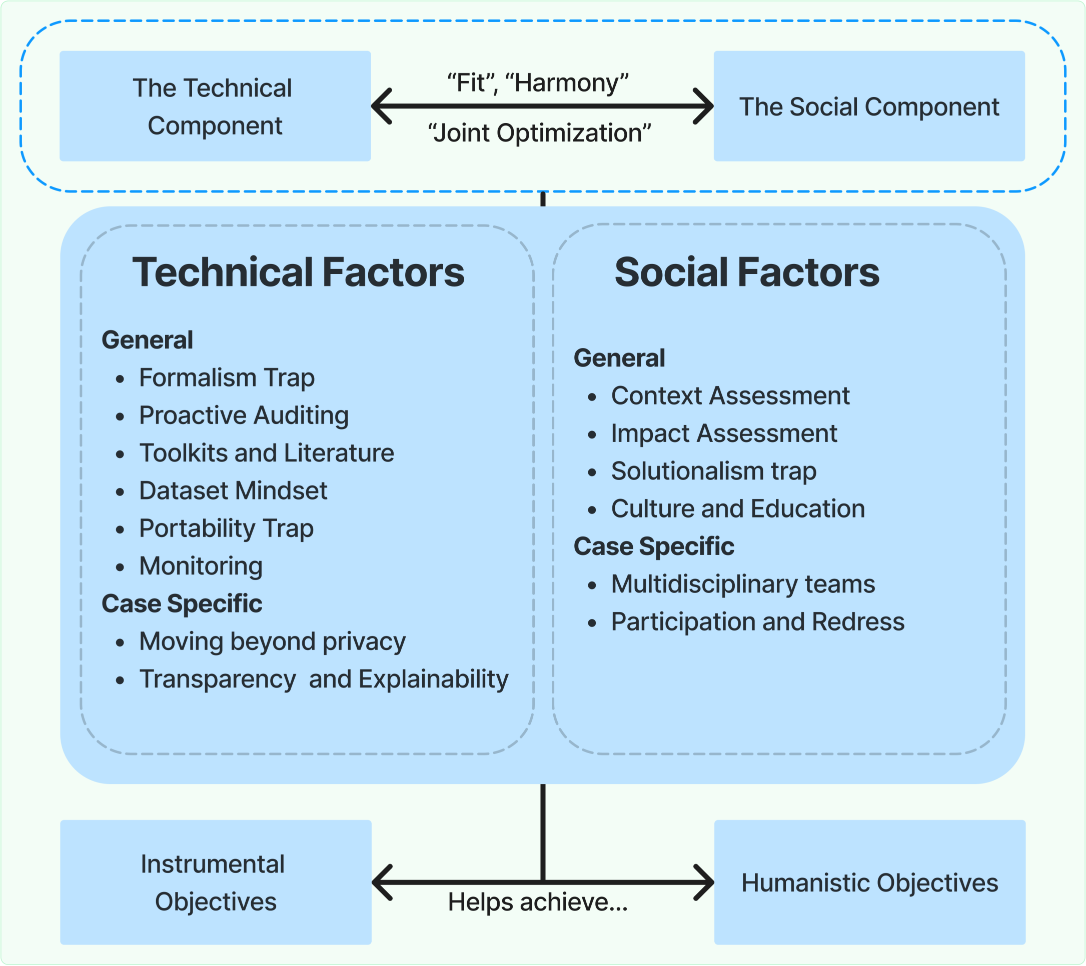


Figure 6.1: Extended Sociotechnical Framework for Algorithmic Fairness.

components that achieve instrumental and humanistic objectives. Through the case study, several factors that contribute to the different objectives were identified, and they have been grouped into four categories. These factors help to achieve instrumental and humanistic objectives for companies.

Instrumental objectives will naturally depend on the context, as different companies have different goals, ownership, and business models. In most cases, the instrumental objectives revolve around reducing bias, promoting equal treatment, and ensuring equitable outcomes, but also ensuring economic profitability and generating business value. Drivers for the instrumental objectives are typically more technical, as they seek to enhance aspects of the algorithms in order to minimize or mitigate discriminatory effects. Examples of these objectives include minimizing the fairness metric so that the algorithm's predictions do not disproportionately disadvantage protected groups or individuals. Another example includes maximizing the accuracy of the model without compromising fairness, ensuring a balance in the fairness-accuracy trade-off, as inaccurate, but fair, predictions can have negative consequences and effects. Differing from ensuring that the statistical measures are in order, objectives revolving around procedural fairness can also be desirable. Allowing individuals affected by the algorithm to both understand the underlying processes and raise concerns where necessary can increase customer satisfaction which can increase revenue. The humanistic objectives typically emphasize the broader social dimensions of algorithmic fairness, recognizing that algorithmic fairness is not solely a technical phenomenon, but a reflection of societal values. Examples of these objectives include the promotion of fairness and justice, such as reducing discrimination and inequality by ensuring that algorithms do not amplify or perpetuate existing bias. The emphasis on creating a more equitable and just society is essential for humanistic objectives. Another example is focusing on enhancing inclusivity and diversity, by ensuring that algorithms are cognizant of the different experiences and perspectives of diverse populations, which can prevent further marginalizing, and instead promote equal opportunity.

As mentioned, the identified factors help achieve instrumental and humanistic objectives. However, it is not given that every category of factors contribute to both type of objectives, in many cases, a category will mostly contribute to one type of objective. Starting with the general technical factors, these factors were found to achieve both types of objectives. Factors such as performing *proactive auditing* in order to avoid bias from the start and having mechanisms in place to handle emerging bias as data and model parameters change are crucial. Using appropriate toolkits can help in properly implementing the technical part of the solution, and ensuring that the outcomes are equitable. They also help achieve certain humanistic objectives, such as *portability.* Recognizing that reusing algorithmic solutions that were designed for a specific context could lead to inaccuracies or cause harm can help prevent algorithmic systems from further marginalization and exclusion, and thus foster both inclusivity and diversity, which are important humanistic objectives. Having a *dataset mindset* is an example of a factor contributing to both objectives. By looking to improve the quality of the dataset, it is easier to achieve both better accuracy and fairness metric, which are instrumental objectives. Having a dataset that better represents the real world can increase diversity, a humanistic objective. Staying up to date with technical solutions, such as the described toolkits is one way that companies can take a more structured and active approach to fairness.

For the case-specific technical factors, these were found to mostly achieve humanistic objectives, such as *transparency and explainability*, which is a factor that can allow for those affected by the algorithmic outcomes to understand underlying processes. One could also argue that this helps achieve instrumental objectives as well, as understanding how the model work can help practitioners discover ways to enhance the algorithm.

The general social factors are mostly concerned with humanistic objectives. Emphasizing the need for improving *culture and education* about fairness both see the broader dimensions of algorithmic fairness. Similarly, performing an *impact assessment* can help understand who is affected by the algorithmic outcomes, and help recognize that algorithmic systems can have significant effects on the life of individuals.

The case-specific social factors were mostly found to have humanistic objectives. *Mechanisms for pooling knowledge across teams* so that one can develop the right solutions depending on the system and context. These factors are case-specific because some companies may only have one machine learning team, such as a small company, and thus sharing knowledge across teams wouldn't be possible. Similarly, companies could, for several reasons, such as privacy or security, not have the possibility to allow affected individuals to participate and raise concerns, even though this would likely be beneficial.

Although the objectives are divided into instrumental and humanistic, these shouldn't be seen as isolated outcomes. Sarker et al. propose a synergistic connection between the instrumental and humanistic objectives, where the positive actions from humanistic outcomes can produce better instrumental outcomes [65]. This synergy was also observed through the interviews, as R5 pointed out that ensuring algorithmic fairness would increase customer trust and satisfaction, which again would lead to more and happier customers and thus increase profits.

The framework has many factors that capture a wide range of considerations, yet there are other potential factors that could have been relevant. One reason why these factors are more relevant is that elements for what constitutes as algorithmic fairness is very context dependent [66], and since interviews do not cover every context where algorithmic fairness is considered, certain factors become more relevant than others.

Chapter 7

Conclusion

After an increase in research on algorithmic fairness, practitioners are now starting to consider algorithmic fairness in their algorithmic decision-making systems. These systems often have a form of human involvement, meaning biases can arise both from human and algorithmic aspects. From a multiple case study of 9 participants from 8 different companies in different industries, this Master's thesis presents how algorithmic fairness is considered and advanced toward by practitioners. This study extends algorithmic fairness research by providing an understanding of Norwegian companies' experiences, approaches, advances, and challenges with algorithmic fairness in algorithmic decision-making systems. By applying a sociotechnical view the study suggests a framework for understanding how harmony between technical and humanistic components achieves both instrumental and humanistic objectives within the context of algorithmic fairness.

The Extended Sociotechnical Framework for Algorithmic Fairness poses several implications for practitioners, providing a simple yet covering illustration of algorithmic fairness. The framework can help practitioners and companies understand how they can approach and advance toward algorithmic fairness, and gain a better understanding of their own situation and context. This can be useful for practitioners, and serve as a motivation to begin with or improve the fairness considerations they are making. For companies, this improved understanding and motivation can help achieve both instrumental and humanistic objectives, by considering both technical and social aspects. For the research community, it provides steps towards gaining knowledge of the different contexts that the literature is being applied. Since the framework is based on projects that use unsanitized, real-world data instead of popular fairness datasets, it offers a different viewpoint from what current literature often provides.

Several limitations are identified for this study. With the study being based on qualitative measures, there is a risk of bias regarding the results and implications, as the author was the only one who attended the interviews. In order to mitigate the risk of misunderstandings and wrongful interpretations, the interviews were transcribed shortly after each interview was conducted, in order to preserve the meanings of the respondents.

Another limitation of the study is the diversity in the considered companies, as they are limited to Norwegian companies or to the Norwegian branches of companies. One discovery was that several legal factors that companies have to consider play a role in how fairness is defined, and some of these laws are not necessarily globally universal. The companies who partook in these interviews were mostly concerned with Norwegian regulations, and while proposals such as the EU AI Act could continue to create more general considerations regarding fair algorithms, there are still regulations that may lead to different perceptions. Factors in the proposed framework were also found in the literature from other countries, but further investigations in different contexts and geographical locations could further improve the reliability of the results from the research.

Future work can expand the framework by finding more relevant factors from industries that were not explored in detail during this research, such as the medical industry, which was found to be a context-specific research area for algorithmic fairness. Another approach could be to create a framework that is only concerned with one particular industry, such as welfare or insurance, and create an improved understanding of algorithmic fairness within that context, which could reduce the risk of falling into the portability trap by acknowledging the context of the algorithmic solution.

Bibliography

[1] Angelika Adensamer, Rita Gsenger and Lukas Daniel Klausner. '"Computer says no": Algorithmic decision support and organisational responsibility'. In: *Journal of responsible technology* 7 (2021), p. 100014.

[2] Alekh Agarwal et al. 'A reductions approach to fair classification'. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 60–69.

[3] Julia Angwin et al. *Machine Bias*. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (visited on 26th May 2022).

[4] Avi Asher-Schapiro. *Global exam grading algorithm under fire for suspected bias*. URL: https://www.reuters.com/article/us-global-tech-education-analysis-trfn-idUSKCN24M29L (visited on 10th May 2022).

[5] Ricardo Baeza-Yates. 'Bias on the web'. In: *Communications of the ACM* 61.6 (2018), pp. 54–61.

[6] Solon Barocas, Moritz Hardt and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. http://www.fairmlbook.org. fairmlbook.org, 2019.

[7] Cynthia Beath et al. 'Expanding the frontiers of information systems research: Introduction to the special issue'. In: *Journal of the Association for Information Systems* 14.4 (2013), p. 4.

[8] Rachel KE Bellamy et al. 'AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias'. In: *IBM Journal of Research and Development* 63.4/5 (2019), pp. 4–1.

[9] Richard Berk et al. 'A convex framework for fair regression'. In: *arXiv preprint arXiv:1706.02409* (2017).

[10] Virginia Braun and Victoria Clarke. 'Using thematic analysis in psychology'. In: *Qualitative research in psychology* 3.2 (2006), pp. 77–101.

[11] Simon Caton and Christian Haas. 'Fairness in machine learning: A survey'. In: *arXiv preprint arXiv:2010.04053* (2020).

[12] Hongyan Chang and Reza Shokri. 'On the privacy risks of algorithmic fairness'. In: *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE. 2021, pp. 292–303.

[13] Irene Chen, Fredrik D Johansson and David Sontag. 'Why is my classifier discriminatory?' In: *Advances in neural information processing systems* 31 (2018).

[14] Alexandra Chouldechova. 'Fair prediction with disparate impact: A study of bias in recidivism prediction instruments'. In: *Big data* 5.2 (2017), pp. 153–163.

[15] Eva Constantaras et al. *Inside the Suspicion Machine.* URL: https://www.wired.com/story/welfare-state-algorithms/ (visited on 26th Mar. 2023).

[16] Kimberlé Crenshaw. 'Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics'. In: *u. Chi. Legal f.* (1989), p. 139.

[17] John W Creswell. *Educational research: Planning, conducting, and evaluating quantitative and qualitative research.* Pearson Education, Inc, 2012.

[18] Daniela S. Cruzes and Tore Dyba. 'Recommended Steps for Thematic Synthesis in Software Engineering'. In: *2011 International Symposium on Empirical Software Engineering and Measurement.* 2011, pp. 275–284. DOI: 10.1109/ESEM.2011.36.

[19] Brian d'Alessandro, Cathy O'Neil and Tom LaGatta. 'Conscientious classification: A data scientist's guide to discrimination-aware classification'. In: *Big data* 5.2 (2017), pp. 120–134.

[20] Jeffrey Dastin. *Amazon scraps secret AI recruiting tool that showed bias against women.* URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (visited on 26th May 2022).

[21] Pablo De Greiff. 'Justice and reparations'. In: *The handbook of reparations* (2006), pp. 451–477.

[22] Mateusz Dolata, Stefan Feuerriegel and Gerhard Schwabe. 'A sociotechnical view of algorithmic fairness'. In: *Information Systems Journal* 32.4 (2022), pp. 754–818.

[23] Pedro Domingos. 'A few useful things to know about machine learning'. In: *Communications of the ACM* 55.10 (2012), pp. 78–87.

[24] Cynthia Dwork. 'Differential privacy'. In: *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33.* Springer. 2006, pp. 1–12.

[25] Lilian Edwards. *Expert explainer: The EU AI Act proposal.* URL: https://www.adalovelaceinstitute.org/wp-content/uploads/2022/04/Expert-explainer-The-EU-AI-Act-11-April-2022.pdf (visited on 18th Apr. 2023).

[26] Marc Eulerich et al. 'A framework for using robotic process automation for audit tasks'. In: *Contemporary Accounting Research* 39.1 (2022), pp. 691–720.

[27] Alessandro Fabris et al. 'Algorithmic Fairness Datasets: the Story so Far'. In: *arXiv preprint arXiv:2202.01711* (2022).

[28] Bela Filep. 'Interview and translation strategies: coping with multilingual settings and data'. In: *Social Geography* 4.1 (2009), pp. 59–70.

[29] Andreas Fuster et al. 'Predictably unequal? The effects of machine learning on credit markets'. In: *The Journal of Finance* 77.1 (2022), pp. 5–47.

[30] Pratyush Garg, John Villasenor and Virginia Foggo. 'Fairness metrics: A comparative analysis'. In: *2020 IEEE International Conference on Big Data (Big Data).* IEEE. 2020, pp. 3662–3666.

[31] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow.* " O'Reilly Media, Inc.", 2022.

[32] GitHub. *GitHub Copilot - Your AI pair programmer.* URL: https://github.com/features/copilot (visited on 22nd May 2022).

[33] Norwegian Government. *Norwegian Position Paper on the European Commission's Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (COM(2021) 206).* URL: https://www.regjeringen.no/contentassets/939c260c81234eae96b6a1a0fd32b6de/norwegian-position-paper-on-the-ecs-proposal-for-a-regulation-of-ai.pdf (visited on 18th Apr. 2023).

[34] Venkat Gudivada, Amy Apon and Junhua Ding. 'Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations'. In: *International Journal on Advances in Software* 10.1 (2017), pp. 1–20.

[35] Christian Haas. 'The Price of Fairness - A Framework to Explore Trade-offs in Algorithmic Fairness'. English (US). In: *40th International Conference on Information Systems, ICIS 2019*. 40th International Conference on Information Systems, ICIS 2019. Publisher Copyright: © 40th International Conference on Information Systems, ICIS 2019. All rights reserved.; 40th International Conference on Information Systems, ICIS 2019 ; Conference date: 15-12-2019 Through 18-12-2019. Association for Information Systems, 2019.

[36] Moritz Hardt, Eric Price and Nathan Srebro. *Equality of Opportunity in Supervised Learning*. 2016. arXiv: 1610.02413 [cs.LG].

[37] Kenneth Holstein et al. 'Improving fairness in machine learning systems: What do industry practitioners need?' In: *Proceedings of the 2019 CHI conference on human factors in computing systems*. 2019, pp. 1–16.

[38] Naja Holten Møller, Irina Shklovski and Thomas T Hildebrandt. 'Shifting concepts of value: Designing algorithmic decision-support systems for public services'. In: *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. 2020, pp. 1–12.

[39] Abigail Z Jacobs. 'Measurement as governance in and for responsible AI'. In: *arXiv preprint arXiv:2109.05658* (2021).

[40] Faisal Kamiran, Toon Calders and Mykola Pechenizkiy. 'Discrimination Aware Decision Tree Learning'. In: *2010 IEEE International Conference on Data Mining*. 2010, pp. 869–874. DOI: 10.1109/ICDM.2010.50.

[41] Michelle E Kiger and Lara Varpio. 'Thematic analysis of qualitative data: AMEE Guide No. 131'. In: *Medical teacher* 42.8 (2020), pp. 846–854.

[42] Barbara Kitchenham. 'Procedures for performing systematic reviews'. In: *Keele, UK, Keele University* 33.2004 (2004), pp. 1–26.

[43] Barbara A Kitchenham, David Budgen and O Pearl Brereton. 'Using mapping studies as the basis for further research–a participant-observer case study'. In: *Information and Software Technology* 53.6 (2011), pp. 638–651.

[44] Jon Kleinberg, Sendhil Mullainathan and Manish Raghavan. *Inherent Trade-Offs in the Fair Determination of Risk Scores*. 2016. arXiv: 1609.05807 [cs.LG].

[45] Vanessa Kohn, Muriel Frank and Roland Holten. 'How Sociotechnical Realignment and Sentiments Concerning Remote Work are Related–Insights from the COVID-19 Pandemic'. In: *Business & Information Systems Engineering* (2023), pp. 1–18.

[46] Nima Kordzadeh and Maryam Ghasemaghaei. 'Algorithmic bias: review, synthesis, and future research directions'. In: *European Journal of Information Systems* 31.3 (2022), pp. 388–409.

[47] Matt J Kusner et al. 'Counterfactual fairness'. In: *Advances in neural information processing systems* 30 (2017).

[48] Max Langenkamp, Allan Costa and Chris Cheung. 'Hiring fairly in the age of algorithms'. In: *arXiv preprint arXiv:2004.07132* (2020).

[49] Li Li et al. 'A review of applications in federated learning'. In: *Computers & Industrial Engineering* 149 (2020), p. 106854.

[50] Anastassia Loukina, Nitin Madnani and Klaus Zechner. 'The many dimensions of algorithmic fairness in educational applications'. In: *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. 2019, pp. 1–10.

[51] Ninareh Mehrabi et al. *A Survey on Bias and Fairness in Machine Learning*. 2022. arXiv: 1908.09635 [cs.LG].

[52] Shira Mitchell et al. 'Algorithmic fairness: Choices, assumptions, and definitions'. In: *Annual Review of Statistics and Its Application* 8 (2021), pp. 141–163.

[53] Lily Morse et al. 'Do the ends justify the means? Variation in the distributive and procedural fairness of machine learning algorithms'. In: *Journal of Business Ethics* (2021), pp. 1–13.

[54] Eduardo Mosqueira-Rey et al. 'Human-in-the-loop machine learning: A state of the art'. In: *Artificial Intelligence Review* 56.4 (2023), pp. 3005–3054.

[55] Briony J Oates. *Researching Information Systems and Computing*. Sage, 2005.

[56] Ziad Obermeyer et al. 'Dissecting racial bias in an algorithm used to manage the health of populations'. In: *Science* 366.6464 (2019), pp. 447–453.

[57] OpenAI. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL].

[58] John W Patty and Elizabeth Maggie Penn. 'Algorithmic fairness and statistical discrimination'. In: *Philosophy Compass* 18.1 (2023), e12891.

[59] Dana Pessach and Erez Shmueli. 'A review on fairness in machine learning'. In: *ACM Computing Surveys (CSUR)* 55.3 (2022), pp. 1–44.

[60] Kai Petersen et al. 'Systematic mapping studies in software engineering'. In: *12th International Conference on Evaluation and Assessment in Software Engineering (EASE) 12*. 2008, pp. 1–10.

[61] Aditya Ramesh et al. *Zero-Shot Text-to-Image Generation*. 2021. arXiv: 2102.12092 [cs.CV].

[62] Sebastian Raschka. 'Model evaluation, model selection, and algorithm selection in machine learning'. In: *arXiv preprint arXiv:1811.12808* (2018).

[63] Drew Roselli, Jeanna Matthews and Nisha Talagala. 'Managing bias in AI'. In: *Companion Proceedings of The 2019 World Wide Web Conference*. 2019, pp. 539–544.

[64] Pedro Saleiro et al. 'Aequitas: A bias and fairness audit toolkit'. In: *arXiv preprint arXiv:1811.05577* (2018).

[65] Suprateek Sarker et al. 'The sociotechnical axis of cohesion for the IS discipline: Its historical legacy and its continued relevance'. In: *MIS quarterly* 43.3 (2019), pp. 695–720.

[66] Andrew D Selbst et al. 'Fairness and abstraction in sociotechnical systems'. In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 59–68.

[67] Mohammad Ahmad Sheikh, Amit Kumar Goel and Tapas Kumar. 'An approach for prediction of loan approval using machine learning algorithm'. In: *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE. 2020, pp. 490–494.

[68] The European Union. *Data protection.* URL: https://commission.europa.eu/law/law-topic/data-protection_en (visited on 18th Apr. 2023).

[69] The European Union. *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS.* URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%5C%3A52021PC0206 (visited on 18th Apr. 2023).

[70] Sahil Verma and Julia Rubin. 'Fairness definitions explained'. In: *Proceedings of the international workshop on software fairness.* 2018, pp. 1–7.

[71] Sahil Verma et al. 'Counterfactual explanations and algorithmic recourses for machine learning: a review'. In: *arXiv preprint arXiv:2010.10596* (2020).

[72] Neil Vigdor. *Apple Card Investigated After Gender Discrimination Complaints.* URL: https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html (visited on 26th May 2022).

[73] Xiaomeng Wang, Yishi Zhang and Ruilin Zhu. 'A brief review on algorithmic fairness'. In: *Management System Engineering* 1.1 (2022), p. 7.

[74] Allison Woodruff et al. 'A qualitative exploration of perceptions of algorithmic fairness'. In: *Proceedings of the 2018 chi conference on human factors in computing systems.* 2018, pp. 1–14.

[75] Julian L Woodward and Raymond Franzen. 'A study of coding reliability'. In: *Public Opinion Quarterly* 12.2 (1948), pp. 253–257.

[76] Fisher Yu et al. 'Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop'. In: *arXiv preprint arXiv:1506.03365* (2015).

Appendix

1. Appendix A: Systematic Mapping Study
2. Appendix B: Sikt Approval and Information Letter
3. Appendix C: Interview Guide

Systematic Mapping Study

Kunnskap for en bedre verden

Department of Computer Science

IT3915 - Master in Informatics, Preparatory Project

Algorithmic fairness: A systematic mapping study

Author:
Fredrik Wilhelm Butler Wang

14th December 2022

Table of Contents

List of Figures

List of Tables

Abstract

Algorithmic fairness is regarded as an emerging research field as automated decision-making is widely used by critical systems. These systems are often found to discriminate and contain bias. Algorithmic fairness technology offers opportunities to mitigate discrimination and improve fairness. This study therefore aims to identify how research on algorithmic fairness is conducted and what types of contribution new research makes. In this study, a mapping of 136 papers from 2018 to 2022 was conducted. A classification schema was developed to categorise recent research on algorithmic fairness based on several classification types. By analysing the classification results, it was discovered that research on algorithmic fairness is increasing, especially towards technical and empirical research. It was also discovered that the definition of fairness varies, which subsequently leads to results that are difficult to compare and that there is a wide range of ways to measure fairness. Datasets employed in the research on algorithmic fairness was found to often be based on the dataset's popularity, regardless of their relevance. Future work can focus on how research on algorithmic fairness defines fairness, and how much of the research is comparable to each other, as well as research on algorithmic fairness within specific domains.

1 Introduction

Algorithmic fairness (AF) has become an increasingly relevant topic in recent years as the use of automatic decision-making and machine learning have evolved to have a significant real-world impact, particularly in areas such as criminal justice, medicine, and finance.

There are several reasons to why these automated systems are used in such a degree. The main benefits include the vast amount of data that algorithms can take into account when making decisions, and the speed at which they can make these decisions compared to humans. The belief that humans will make biased and subjective decisions as opposed to algorithms is also a common reason. However algorithms are found to contain biases, even though when actions to ensure fairness is believed to be in place (Corbett-Davies et al. 2017). A reason for this bias stems from the data used to train algorithms contain bias themselves. The bias in algorithms often goes unnoticed because algorithms are often either viewed in the same way that big data is viewed, namely objective, and accurate (Boyd and Crawford 2012), or because they are too difficult to understand for the casual user, and even difficult to interpret by experts (Wong 2020).

The possible damage from algorithmic bias can be enormous, and AF should therefore be considered important. Concerns for AF has made headlines in recent years after several discoveries. In the criminal justice field, it has been revealed that an algorithm used by the United States system falsely predicted future criminal behavior among African-Americans at twice the rate as it did for white people (Angwin et al. 2016). Interestingly enough, developers of this system claimed that a fairness mechanism was implemented in the algorithm (Dieterich et al. 2016). Researchers and developers have proposed methods to the detection and mitigation of bias in algorithms. Yet there is several critics to these proposed methods, primarily because the focus lies on implementing fairness perspectives into the algorithms, and solving it as a technical problem (Wong 2020), Dolata et al. 2022). Previous research also indicates that AF should be pursued by stakeholders in a business perspective, both for legal, branding and user-trust reasons (Woodruff et al. 2018). To mitigate this bias and ensure fairness is therefore a difficult task, but a desired goal.

This article aims to explore the current state of research on algorithmic fairness. The objective of this mapping study is to give an updated view on algorithmic fairness research and identify research trends.

The research objective leads to the following research questions:

- RQ1: How has research on Algorithmic Fairness changed between 2018 and 2022?
- RQ2: How do technical frameworks achieve Algorithmic Fairness?

In this article, I present results from a mapping study on research on algorithmic fairness from 2018 to 2022. This is done with a search strategy similar to a previous literature review, that focused on studies from 2017 to 2020 (Dolata et al. 2022). 136 relevant papers were found during the time span of 2018 to 2022. In order to address RQ1, papers were mapped according to several classification types. With RQ2, I look at the approaches frameworks take in order to combat algorithmic bias and unfairness.

The paper proceeds as follows: Section 2 introduces the background of the study and other conducted studies of relevance. Section 3 presents the research method. Section 4 shows the results and visualises the findings from the mapping study. Section 5 discusses the findings and the relation to the research questions, and concludes with recommendations for future work.

2 Background

2.1 Fairness

Fairness is a broad and complicated research field, and there is no universal definition of fairness (Mehrabi et al. 2021), which again makes it difficult to solve fairness related problems. In a broad sense, fairness can be seen as the absence of discrimination and the presence of impartiality, but how fairness is perceived is often affected by culture and personal preferences. Achieving fairness is often desired in society, but difficult to achieve in practice. One can therefore use and propose different definitions of fairness in order to address algorithmic unfairness (Mehrabi et al. 2021). Typically on would distinguish between group and individual fairness, and use different metrics to measure this fairness (Pessach and Shmueli 2020).

2.2 Algorithmic fairness

Algorithmic fairness aims to detect and mitigate harm or benefits given to subgroups by automated decision-making. It's relevance stems from the idea that algorithms may inherently be biased, based on the historical biases that have been learned and kept. AF can be viewed from different perspectives, such as technical (Lepri et al. 2018), philosophical (Binns 2018) and sociotechnical (Dolata et al. 2022). Addressing AF is important because an increasing amount of automated decisions affect peoples life, such as job application processes (Noble et al. 2021), bank loans (Mukerjee et al. 2002), or penalty determination in justice systems (Završnik 2020). AF is a relatively new research field, that has received an increasing amount of interest in the last couple of years

(Mehrabi et al. 2021). Mitigating algorithmic unfairness is not a trivial task because an increase of fairness could lead to a decrease of accuracy, resulting in a accuracy-fairness trade-off that needs to be considered by stakeholders using the algorithms (Wang et al. 2022, Pessach and Shmueli 2020).

The term algorithmic bias is also used to describe the potential biases and unfairness that can arise in algorithms and machine learning systems. The two terms are periodically used interchangeable, but there is a difference between the two terms. AF refers to the idea that algorithms should be fair and unbiased, while algorithmic bias refers to the inherent biases that can arise in algorithms due to the data and assumptions used to train them.

2.3 Fairness framework

A fairness framework refers to a set of tools, standards and conventions that provide a foundation for tackling fairness in a technical solution. In the context of AF, it includes methods for mitigating, detecting and handling AF. Examples include the toolkit *AI Fairness 360* (IBM Research 2022), and simple conceptual ones used in research (Kleinberg et al. 2018)

2.4 Fairness metric

A fairness metric is a metric that quantify and statistically measure fairness in algorithms. A metric for fairness can vary depending on the notion of fairness.

2.5 Algorithmic fairness datasets

These datasets are some of the most commonly used datasets in the literature of algorithmic fairness. What they have in common is that they have one or more sensitive variables, that is a variable that can be used do discriminate against certain groups or individuals. Sensitive variables typically contain personal characteristics or demographic information that are protected by law, such as race, gender, age, sexual orientation, or religion. The datasets stem from different sources and research fields, although it should be noted that most of the datasets found in the literature originate from the west (Mehrabi et al. 2021).

Popular datasets include: Adult - based on a US population survey that include socially relevant data. COMPAS - created as an external audit of racial biases. German Credit - containing loan applicants.

2.6 Existing literature review

AF is an emerging research field that has become even more important as the number of decisions made by algorithms is increasing. Since AF is a subgroup of the fairness term it can be part of reviews that cover overall fairness, either in a broad context, or in a specific research field, such as machine learning or facial detection systems. It can also be covered in reviews limited to AF only.

Dolata et al. 2022 Covered studies from 2017 and 2020 and presented AF as a sociotechnical concept. A systematic analysis of 310 articles is performed. They argue that algorithmic unfairness does not arrive solely from algorithms or data, but also from the adaptations between technological and social components. The papers suggests that research on AF is focused on solving issues through technical solutions. However to solve issues where both social and technical components are involved, the authors argued that the entire sociotechnical system should be addressed.

Kordzadeh and Ghasemaghaei 2022 Performed a systematic review of 56 relevant papers between 2010-2019, most of them conceptual. The study concludes that non-technical studies focuses on conceptual solutions, and that there is little empirical research that highlights the effects of AF from a non-technical perspective.

3 Methodology

3.1 Mapping procedure

In the early phase, searches where performed in Scopus to test what results different search strings would yield. This was done in order to get a grasp of what keywords could give the most relevant searches. Dolata et al. 2022 provides a systematic analysis of 310 algorithmic fairness articles from 2017 to 2020. In relevance to RQ1, it was decided that it would be favorable to reproduce this search, but for the time period of 2018-2022, to see how research on AF has changed since Dolata et al. 2022 was conducted.

3.2 Data sources and search strategy

To perform a systematic search, Scopus was used. Scopus can handle complex search strings, has several options for filtering results, and has been used in previous mapping studies (Kordzadeh and Ghasemaghaei 2022) and systematic analyses (Dolata et al. 2022) of AF.

The final search string in the Scopus syntax is thus, as it is in Dolata et al. 2022:

((”*fair*” PRE/1 (”ML” OR ”machine learn-

ing" OR "AI" OR "artificial intelligence"))
OR (("algorithmic*" OR "AI" OR "ML" OR
"machine learning" OR "artificial intelligence")
PRE/1 ("fair*" OR "justi*" OR "bias*" OR
"unfair*")))

This is a broad search query, that can capture a substantial amount of possibly relevant literature.

3.3 Study selection

The process of study selection is shown in Figure 1, with the number of papers for each step. In Scopus, the search string returned 988 results. Results where then limited to studies between 2018-2022, subject areas of *Computer Science* and *Business, Management and Accounting*, document types of *Articles* and *Conference papers*, source type as *Journals*, as well as only papers written in *English.*

With the results limited as described, Each paper's title and abstract was read and marked as relevant or not. A set of inclusion and exclusion criteria where developed, the final criteria are as follows:

Inclusion:

1. Discusses fairness related to a technical solution
2. Makes a contribution to fairness or discusses fairness as a core concept

Related to a technical solution here means that a paper doesn't need to be technical itself, but it has to contain research on something that is, so a paper discussing whether or not the US justice system is unfair would be excluded.

Exclusion:

1. No individual contribution (editorials, commentaries, calls for papers, or tutorials)
2. Words in query not used in the intended meaning
3. Refers to systematic deviation and not actual unfair treatment
4. Only refers to unfairness in general terms, no link between technology and discrimination
5. Only refers to fairness in the future work section or as a motivation

3.4 Data extraction

After the study selection process, a classification schema was defined (Table 5). Based on the title and abstract the papers were categorized based on the attributes *Type, Overall Context, Context, Empirical or Conceptual, Fairness Focus*, and *Output.* All attributes were used to map how research on AF is conducted, whereas *Fairness focus* was also used to determine if the papers only focuses on AF or if they take a broader approach, and also to determine what term is used for AF. As mentioned in 2.1 and 2.2 the definition of fairness and AF varies in the literature and this classification category highlights this challenge. Regarding RQ2, all technical frameworks where also categorized based on their focus area. Studies were categorized in a tabular form (i.e. Excel) to ensure that comparison and analysis of the results could be fluently conducted and visualised. Appendix B contains the full classification of all relevant papers.

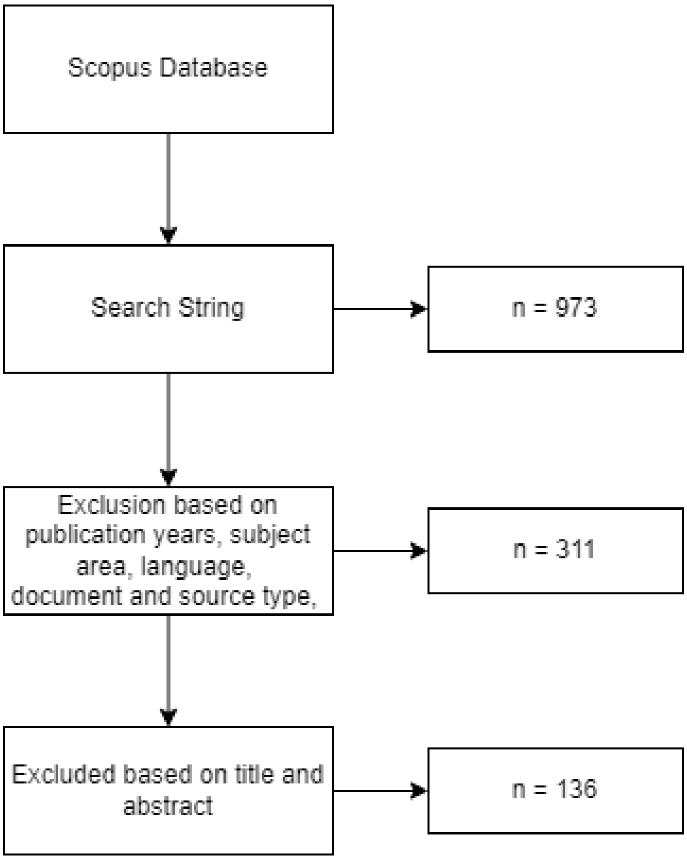

Scopus Database
Search String
n = 973
Exclusion based on publication years, subject area, language, document and source type,
n = 311
Excluded based on title and abstract
n = 136

Figure 1: The study selection process

4 Findings

This section presents the data extracted from the literature review. The section is further divided into subsections for each research question for a clearer visualisation and highlighting of the findings from the 136 papers.

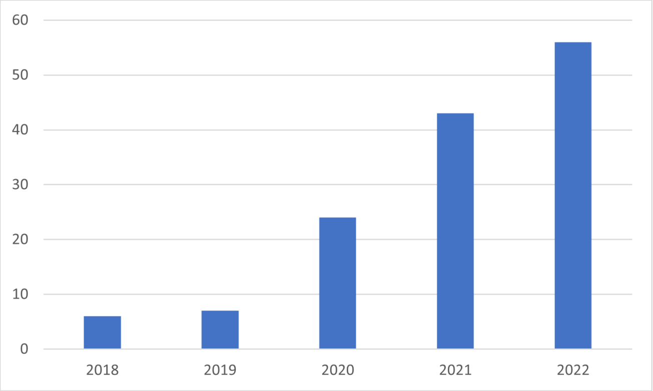


Figure 2: Publication frequency, 2018-2022.

4.1 RQ1: How has research on Algorithmic Fairness changed between 2018 and 2022?

Figure 2 shows the publication frequency related to AF given by the inclusion and exclusion criteria, resulting in a total of 136 papers. It is observed that the frequency of papers is increasing each year, with the year with the highest numbers of papers being 2022.

The number of technical papers, 91, is higher than the number of non-technical papers, 45. This is also holds for each year, except for 2018, when there was 3 papers in each category (Figure 3).

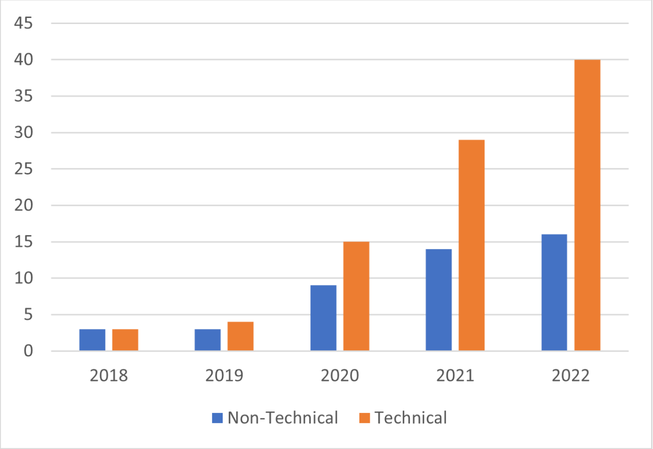


Figure 3: Technical and non-technical papers, 2018-2022.

Empirical research is more common than conceptual research in the field of AF (Figure 4). This difference is more evident in technical papers, where 84.6 percent of research is empirical, as to 57.8 percent for non-technical papers.

The domain of the research is dominated by a generic domain, except from 2019, although this is a small sample size (Figure 5). For the papers written in domain-specific contexts, the sample size is again small for 2018-2020, but for 2021 and 2022, one can observe that economical and medical papers are the most dominant ones (Figure 6).

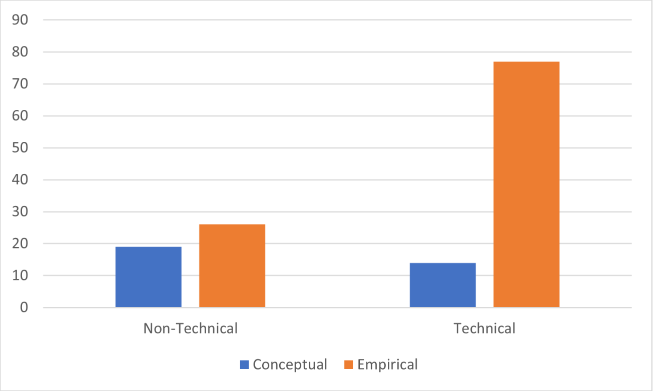


Figure 4: Number of technical and non-technical papers given type of research.

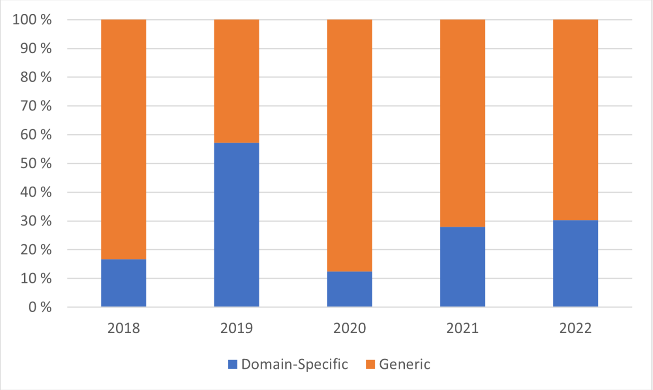


Figure 5: Percentage of Generic and Domain-specific papers, 2018-2022.

Each paper was classified based on their fairness focus. This attribute can be used to indicate how much pertinence a paper has to AF. 62.5 percent of the papers had a main focus on AF, if one sees a focus on algorithmic bias as also being of high pertinence, then 70,7 percent of the papers have a high pertinence. The remaining papers either have a focus on general fairness that also includes AF, or they focus on achieving or measuring fairness in algorithms without explicitly referring to AF (Ferry et al. 2022). The full distribution of fairness focus is shown in table 1

38 (27.9 percent) of the papers either present their results or have results as a core part of their output. The full distribution of outputs is shown in table 2. A more detailed description of the different outputs is given in Table 6. It is observed that results are often based on industry-known datasets, such as those described in subsection 2.5.

4.2 RQ2: How do technical frameworks achieve Algorithmic Fairness?

As shown in figure 3, the literature is dominated by technical perspectives, Dolata et al. 2022 also notes this. All

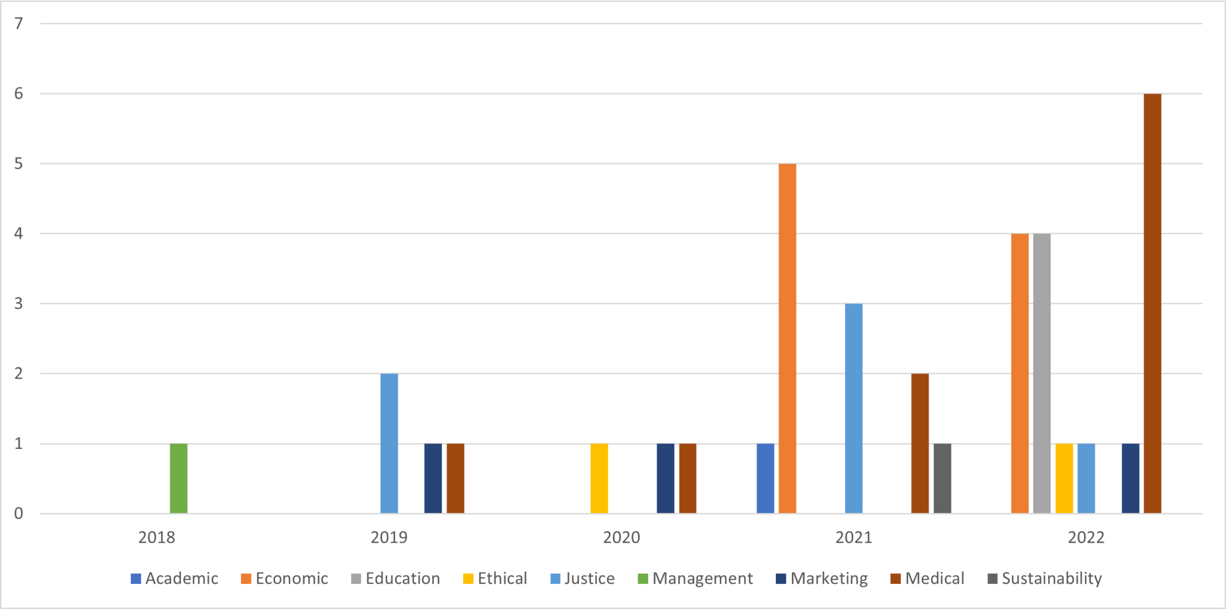


Figure 6: Domain-specific research, 2018-2022.

Fairness focus	Frequency
Algorithmic fairness	85
Algorithmic bias	11
General	34
Bias Mitigation	1
Dataset Bias	1
Detection	1
Mitigation	1
Social Influence Bias	1
Statistical Fairness	1

Table 1: Fairness focus, 2018-2022

Output	Frequency
Algorithm and Results	9
Framework	43
Framework and Metric	15
Framework and Results	1
Metric	8
Overview	32
Results	28

Table 2: Outputs, 2018-2022

technical papers that where classified has having a framework as part of the output (the contribution type of the paper), were mapped based on the area that the framework focuses on. In total, 43.4 percent of all papers had a framework as the output, and 59.3 percent of all technical papers had framework as part of their output. The distribution of the focus area of the frameworks is shown in table 3, and the full classification of the technical frameworks are shown in appendix A. The distribution shows that classification tasks is the most common area that frameworks focuses on. Classification is a common and broad task in machine learning and the categorisation aspect that it involves is strongly related typical automated decision-making, where the system has a certain options to choose from when making a decision. Classification as the main focus area of frameworks is also reported by Mehrabi et al. 2021.

Besides the area that the fairness mitigation focuses on, it is also observed that despite the context of the papers, popular fairness datasets such as COMPAS and German Credit are used to measure and report results.

In the relevant papers, it is observed that when a method for mitigating algorithmic unfairness is proposed, a definition of fairness is often also proposed. Several definitions that constitute what achieving fairness involves is therefore used within the relevant papers. There are also those frameworks that let researchers make use of different definitions and metrics for evaluating fairness (Wexler et al. 2019, Bellamy et al. 2019). However these papers do not make up the majority.

Table 4 presents a summary of the main findings contributing to addressing the research questions: (RQ1): *How has research on Algorithmic Fairness changed between*

Focus Area	Frequency
Adversarial learning	3
Auto-encoding	1
Casual inference	2
Classification	35
Clustering	1
Language model	1
Non-specific	4
Recommender systems	3
Regression	2
Word embedding	1

Table 3: Focus areas of technical frameworks that combat algorithmic bias and unfairness

2018 and 2022? (RQ2): *How do technical frameworks achieve Algorithmic Fairness?*

5 Discussion

In this paper, a mapping study of 136 relevant papers have been conducted, to observe how research on algorithmic fairness has evolved and also gained traction in the research community. The different classification types makes it possible to identify change in research direction for the last 5 years. One a general note, it is also recommended for future reviews on AF to perform workshops etc. to decide on and define the different categories for classifying research on AF. This is a difficult but important task as it is necessary in order to draw useful observations and conclusions from the relevant papers.

5.1 RQ1: How has research on Algorithmic Fairness changed between 2018 and 2022?

Technical research was more common than non-technical research. An explanation for why this is the case could be from the study selections, where the subject areas *Computer Science* and *Business, Management and Accounting* where chosen, as well as the inclusion criteria that states that even tough a paper can be non-technical, it's topic should be related to a technical solution, and as a consequence of this, more technical papers could have been selected. In addition, the dominance of technical papers could be inherent from the definition of AF, as the algorithmic and automated decision-making aspects are technical by nature. Other literature also find that most research view AF as a technical discipline (Dolata et al. 2022). An explanation of this could be that AF aims to use mathematical methods to measure bias in machine learning and to use these metrics to reduce discrimination against subgroups, which again leads to unfairness being viewed as a technical issue that decision-making systems can solve through technical approaches.

It was clear that empirical research was more common than conceptual research. There are several possible reasons for why this was the case. Empirical research on AF could be easier to conduct and publish than conceptual research. This is because empirical research would typically involve collecting data and analyzing it using statistical methods, while conceptual research often involves developing new theories that are more difficult to test and validate. Given that there exists a large amount of fairness datasets that researches can use to compare their results against, it is easier to get results, which means that empirical research could be faster to conduct.

Empirical research is often more directly applicable to practical problems and can provide more concrete and specific results. This can make it more attractive to researchers (and journals), as the results of empirical research can be more easily translated into practical applications. Even tough their is little indication from the literature that results and frameworks from researchers can be instantly applied in industry use, it is still possible to achieve. On the other hand, pure conceptual research could be harder to translate into results.

Conceptual research is often more abstract and theoretical, and may require a deeper understanding of AF in order to be properly evaluated. This could imply that it is more challenging to conduct and publish conceptual research, as it may require more specialized knowledge and expertise. Given that research on AF has gained traction in recent years it could mean that the required knowledge for conceptual research is limited and lead to less research in this direction.

The fairness focus in the analysed papers was for the majority centered around AF itself, and only 25 percent had a general focus on fairness. Combined with a relatively large amount of papers being included in this study, the fairness focus distribution resonates well with the observation that AF, which is one of several fairness-related research fields, is being specifically researched, and not only being looked at in a broader sense. (Kordzadeh and Ghasemaghaei 2022) points out that other literature reviews on AF or algorithmic bias have limitations in either being to broad, such as focusing on data science ethics, or being too narrow, such as only looking into fairness and bias in facial analysis systems. The results from this study indicates that there is enough literature for conducting reviews on AF as the sole topic.

Most studies have a general domain, and economical and medical are the only domain-specific fields that receive a lot of attention. Given that AF plays an increasingly important role in several aspects of peoples lives through automated decision-making, I argue that more research should be provided within specific domains, as it can ensure that fairness measures are taken in all fields that utilize automated decision-making.

Research Question	Findings
RQ1	Research on AF has seen an increase, with a shift towards technical and empirical research in a generic context, where medical and economic fields make up the majority of domain specific research. Technical research has a focus on frameworks and metrics. Non-technical research tend to focus on overviews and research suggestions, with less consideration given to frameworks and metrics.
RQ2	Frameworks are the most popular type of contribution for research on AF. Classification, a very broad topic, is the most common focus area for frameworks in achieving AF. Research that produce frameworks often report their results based on their own definition of fairness. Results are often based on the same popular fairness datasets that may not be best suited for the specific task at hand.

Table 4: Summary of results

I also observe that most research conducted use the same group of fairness datasets, such as Adult, COMPAS and German Credit. The use of these datasets is not because of their quality, but instead from their popularity and use cases in influential articles (Angwin et al. 2016). I argue that datasets in research instead should be chosen by their relevance, based on the domain, task and role for the problem at hand, using recommendations such as those found in Fabris et al. 2022.

5.2 RQ2: How do technical frameworks achieve Algorithmic Fairness?

Papers with technical perspective often share the premise that constraints that promote equality and justice are put on the algorithm, and that the goal is to maximise accuracy whilst being subject to these constraints. Subsequently, these approaches often employ a technical solution to the problem, where a solution is good if the problem can be solved. The fairness constraint can vary, and therefore the achieved fairness is not automatically comparable to results from other frameworks. As an effect of this, I argue that the varying definition of what achieving fairness means, further divides AF research, and ultimately makes it easier to achieve results through research, because one has more control over the fairness definition, but more difficult to compare these results with each other.

The majority of frameworks focused on mitigating unfairness within the context of classification. This is not an unexpected result, as the same has been reported by others (Mehrabi et al. 2021). Automated decision-making systems often perform classifications, where data is inputted and then categorised by the system. Examples of this is loan applications, where the predicted value would either be to grant a loan or to not grant a loan. It is also a very broad topic, with tasks such as Binary classification, multi-class classification, multi-label classification and imbalanced classification. There are also several algorithms that can perform these type of tasks, such as logistic regression, naive Bayes and support vector machines to name a few.

For future work I recommend a focus on possible solutions to how frameworks and metrics for algorithmic fairness can be standardized for research and industry use. This would allow for research that is done on AF to be more efficiently used in real-world systems, which are known to suffer from algorithmic unfairness. Similarly, to research and develop frameworks that can accept multiple definitions of fairness would sort out issues related to incompatibility. Some of the relevant papers take an approach where this is possible, but a common denominator for the papers is that they give their own definition of fairness. Being able to test a framework against different fairness definitions and metrics could also improve the resilience of the proposed frameworks, as researchers wouldn't be able to tweak the definition of fairness based on what gives the best results.

Bibliography

Angwin, J., J. Larson and L. Kirchner (2016). *Machine Bias.* URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (visited on 8th Dec. 2022).

Bellamy, Rachel KE et al. (2019). 'AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias'. In: *IBM Journal of Research and Development* 63.4/5, pp. 4–1.

Binns, Reuben (2018). 'What can political philosophy teach us about algorithmic fairness?' In: *IEEE Security & Privacy* 16.3, pp. 73–80.

Boyd, Danah and Kate Crawford (2012). 'Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon'. In: *Information, communication & society* 15.5, pp. 662–679.

Corbett-Davies, Sam et al. (2017). 'Algorithmic decision making and the cost of fairness'. In: *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pp. 797–806.

Dieterich, William, Christina Mendoza and Tim Brennan (2016). 'COMPAS risk scales: Demonstrating accuracy equity and predictive parity'. In: *Northpointe Inc* 7.4.

Dolata, Mateusz, Stefan Feuerriegel and Gerhard Schwabe (2022). 'A sociotechnical view of algorithmic fairness'. In: *Information Systems Journal* 32.4, pp. 754–818. DOI: https://doi.org/10.1111/isj.12370. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/isj.12370. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/isj.12370.

Fabris, Alessandro et al. (2022). 'Algorithmic Fairness Datasets: the Story so Far'. In: *arXiv preprint arXiv:2202.01711*.

Ferry, Julien et al. (2022). 'Improving fairness generalization through a sample-robust optimization method'. In: *Machine Learning*, pp. 1–62.

IBM Research (2022). *AI Fairness 360.* URL: http://aif360.mybluemix.net/ (visited on 9th Dec. 2022).

Kleinberg, Jon et al. (2018). 'Algorithmic fairness'. In: *Aea papers and proceedings.* Vol. 108, pp. 22–27.

Kordzadeh, Nima and Maryam Ghasemaghaei (2022). 'Algorithmic bias: review, synthesis, and future research directions'. In: *European Journal of Information Systems* 31.3, pp. 388–409. DOI: 10.1080/0960085X.2021.1927212. eprint: https://doi.org/10.1080/0960085X.2021.1927212. URL: https://doi.org/10.1080/0960085X.2021.1927212.

Lepri, Bruno et al. (2018). 'Fair, transparent, and accountable algorithmic decision-making processes'. In: *Philosophy & Technology* 31.4, pp. 611–627.

Mehrabi, Ninareh et al. (July 2021). 'A Survey on Bias and Fairness in Machine Learning'. In: *ACM Comput. Surv.* 54.6. ISSN: 0360-0300. DOI: 10.1145/3457607. URL: https://doi.org/10.1145/3457607.

Mukerjee, Amitabha et al. (2002). 'Multi–objective evolutionary algorithms for the risk–return trade–off in bank loan management'. In: *International Transactions in operational research* 9.5, pp. 583–597.

Noble, Sean M, Lori L Foster and S Bartholomew Craig (2021). 'The procedural and interpersonal justice of automated application and resume screening'. In: *International Journal of Selection and Assessment* 29.2, pp. 139–153.

Pessach, Dana and Erez Shmueli (2020). 'Algorithmic fairness'. In: *arXiv preprint arXiv:2001.09784*.

Wang, Xiaomeng, Yishi Zhang and Ruilin Zhu (2022). 'A brief review on algorithmic fairness'. In: *Management System Engineering* 1.1, pp. 1–13.

Wexler, James et al. (2019). 'The what-if tool: Interactive probing of machine learning models'. In: *IEEE transactions on visualization and computer graphics* 26.1, pp. 56–65.

Wong, Pak-Hang (2020). 'Democratizing algorithmic fairness'. In: *Philosophy & Technology* 33.2, pp. 225–244.

Woodruff, Allison et al. (2018). 'A qualitative exploration of perceptions of algorithmic fairness'. In: *Proceedings of the 2018 chi conference on human factors in computing systems*, pp. 1–14.

Završnik, Aleš (2020). 'Criminal justice, artificial intelligence systems, and human rights'. In: *ERA Forum.* Vol. 20. 4. Springer, pp. 567–583.

Appendix

The appendix presents the classifications for the frameworks focus area (A), and the classification schema (B), which includes the classification of each paper. Tables 5 and 6 give a more detailed description of the attributes in the classification schema.

A Classification Descriptions

Classification type	Description
Type	Technical or Non-Technical
Overall Context	Generic or Domain-Specific
Context	Generic or Specific (with specified context)
Research	Empirical or conceptual research
Fairness focus	The aspect/type in which fairness is discussed. Articles where typically found to focus on AF, algorithmic bias, fairness in broader terms with AF as part of the study or a more narrow approach such as bias mitigation
Output	What does the paper produce and contribute with, described in Table 6

Table 5: Classification type

Output type

Contribution type	Description
Metric	New metric(s) for evaluating and measuring fairness in a technical solution
Framework	A new method for achieving fairness
Algorithm	Similar to framework, generally more theoretical
Overview	A review of different studies and potentially proposals for future research or research directions
Results	Paper is task oriented and presents results from the task

Table 6: Output type

B Analysed papers

A Framework focus areas

Paper	Focus Area
Wexler J., Pushkarna M., Bolukbasi T., Wattenberg M., Viegas F., Wilson J. (2020). The what-if tool: Interactive probing of machine learning models	Non-specfic
Bellamy R.K.E., Mojsilovic A., Nagar S., Ramamurthy K.N., Richards J., Saha D., Sattigeri P., Singh M., Varshney K.R., Zhang Y., Dey K., Hind M., Hoffman S.C., Houde S., Kannan K., Lohia P., Martino J., Mehta S. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias	Classification
Paulus J.K., Kent D.M. (2020). Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities	Non-specific
Gu X., Angelov P.P., Soares E.A. (2020). A self-adaptive synthetic over-sampling technique for imbalanced classification	Classification
Edizel B., Bonchi F., Hajian S., Panisson A., Tassa T. (2020). FaiRecSys: mitigating algorithmic bias in recommender systems	Recommender systems
Altman M., Wood A., Vayena E. (2018). A Harm-Reduction Framework for Algorithmic Fairness	Casual inference
Zehlike M., Hacker P., Wiedemann E. (2020). Matching code and law: achieving algorithmic fairness with optimal transport	Classification
Checco A., Bracciale L., Loreti P., Pinfield S., Bianchi G. (2021). AI-assisted peer review	Word embedding
Fu R., Huang Y., Singh P.V. (2021). Crowds, lending, machine, and bias	Classification
Lyu L., Li Y., Nandakumar K., Yu J., Ma X. (2022). How to Democratise and Protect AI: Fair and Differentially Private Decentralised Deep Learning	Adversarial
Valdivia A., Sánchez-Monedero J., Casillas J. (2021). How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness	Classification
Grari V., Ruf B., Lamprier S., Detyniecki M. (2020). Achieving Fairness with Decision Trees: An Adversarial Approach	Classification
Fitzsimons J., Al Ali A., Osborne M., Roberts S. (2019). A general framework for fair regression	Regression
Ashokan A., Haas C. (2021). Fairness metrics and bias mitigation strategies for rating predictions	Recommender systems
Wang Q., Xu Z., Chen Z., Wang Y., Liu S., Qu H. (2021). Visual analysis of discrimination in machine learning	Classification
Salazar R., Neutatz F., Abedjan Z. (2021). Automated feature engineering for algorithmic fairness	Classification
Oneto L., Donini M., Pontil M., Shawe-Taylor J. (2020). Randomized learning and generalization of fair and private classifiers: From PAC-Bayes to stability and differential privacy	Classification
Yoon T., Lee J., Lee W. (2020). Joint Transfer of Model Knowledge and Fairness over Domains Using Wasserstein Distance	Classification
Varley M., Belle V. (2021). Fairness in machine learning with tractable models	Casual inference
Zhang T., Zhu T., Gao K., Zhou W., Yu P.S. (2021). Balancing Learning Model Privacy, Fairness, and Accuracy With Early Stopping Criteria	Classification
Morse L., Teodorescu M.H.M., Awwad Y., Kane G.C. (2021). Do the Ends Justify the Means? Variation in the Distributive and Procedural Fairness of Machine Learning Algorithms	Non-Specific
Kehrenberg T., Chen Z., Quadrianto N. (2020). Tuning Fairness by Balancing Target Labels	Classification
Zhang K., Khosravi B., Vahdati S., Faghani S., Nugen F., Rassoulinejad-Mousavi S.M., Moassefi M., Jagtap J.M. M., Singh Y., Rouzrokh P., Erickson B.J. (2022). Mitigating Bias in Radiology Machine Learning: 2. Model Development	Classification
von Zahn M., Feuerriegel S., Kuehl N. (2022). The Cost of Fairness in AI: Evidence from E-Commerce	Classification
Akter S., Dwivedi Y.K., Sajib S., Biswas K., Bandara R.J., Michael K. (2022). Algorithmic bias in machine learning-based marketing models	Classification
Li J., Moskovitch Y., Jagadish H.V. (2021). Denouncer: Detection of unfairness in classifiers	Classification
Oneto L. (2020). Learning fair models and representations	Non-Specific
Wang M., Zhang Y., Deng W. (2022). Meta Balanced Network for Fair Face Recognition	Classification
Franco D., Navarin N., Donini M., Anguita D., Oneto L. (2022). Deep fair models for complex data: Graphs labeling and explainable face recognition	Classification
Mengesha Z., Heldreth C., Lahav M., Sublewski J., Tuennerman E. (2021). “I don’t Think These Devices are Very Culturally Sensitive.”—Impact of Automated Speech Recognition Errors on African Americans	Language model
Ahmed Z., Vidgen B., Hale S.A. (2022). Tackling racial bias in automated online hate detection: Towards fair and accurate detection of hateful users with geometric deep learning	Classification
Anahideh H., Asudeh A., Thirumuruganathan S. (2022). Fair active learning	Classification

Paper	Focus Area
Petrović A., Nikolić M., Radovanović S., Delibašić B., Jovanović M. (2022). FAIR: Fair adversarial instance re-weighting	Adversarial
Pessach D., Shmueli E. (2021). Improving fairness of artificial intelligence algorithms in Privileged-Group Selection Bias data settings	Classification
Radovanović S., Petrović A., Delibašić B., Suknović M. (2021). A fair classifier chain for multi-label bank marketing strategy classification	Classification
Plečko D., Meinshausen N. (2020). Fair data adaptation with quantile preservation	Classification
Castelnovo A., Cosentini A., Malandri L., Mercorio F., Mezzanzanica M. (2022). FFTree: A flexible tree to handle multiple fairness criteria	Classification
Sokol K., Santos-Rodriguez R., Flach P. (2022). FAT Forensics: A Python toolbox for algorithmic fairness, accountability and transparency[Formula presented]	Classification
Kim K., Ohn I., Kim S., Kim Y. (2022). SLIDE: A surrogate fairness constraint to ensure fairness consistency	Classification
Folorunso S., Ogundepo E., Basajja M., Awotunde J., Kawu A., Oladipo F., Abdullahi I. (2022). FAIR Machine Learning Model Pipeline Implementation of COVID-19 Data	Clustering
Risser L., Sanz A.G., Vincenot Q., Loubes J.-M. (2022). Tackling Algorithmic Bias in Neural-Network Classifiers using Wasserstein-2 Regularization	Classification
Liu S., Vicente L.N. (2022). Accuracy and fairness trade-offs in machine learning: a stochastic multi-objective approach	Classification
Li C., Xing W., Leite W. (2022). Using fair AI to predict students' math learning outcomes in an online platform	Regression
Sha L., Rakovic M., Das A., Gasevic D., Chen G. (2022). Leveraging Class Balancing Techniques to Alleviate Algorithmic Bias for Predictive Tasks in Education	Classification
Liu S., Sun S., Zhao J. (2022). Fair Transfer Learning with Factor Variational Auto-Encoder	Auto-encoding
Kairouz P., Liao J., Huang C., Vyas M., Welfert M., Sankar L. (2022). Generating Fair Universal Representations Using Adversarial Models	Adversarial
Andreeva O., Almeida M., Ding W., Crouter S.E., Chen P. (2022). Maximizing Fairness in Deep Neural Networks via Mode Connectivity	Classification
Zhao H., Gordon G.J. (2022). Inherent Tradeoffs in Learning Fair Representations	Classification
Nguyen D., Gupta S., Rana S., Shilton A., Venkatesh S. (2021). Fairness improvement for black-box classifiers with Gaussian process	Classification
Vermeer N., Boer A., Winkels R. (2021). Survivorship bias mitigation in a recidivism prediction tool	Classification
Kanamori K., Arimura H. (2021). Fairness-aware decision tree editing based on mixedinteger linear optimization	Classification
Popa A. (2021). Fairness Embedded Adaptive Recommender System: A Conceptual Framework	Recommender systems

B Complete analysis of papers

Paper	Type	Overall Context	Context	Research	Fairness focus	Output
Lee M.K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management	Technical	Domain-Specific	Management	Empirical	AF	Results
Lambrecht A., Tucker C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of stem career ads	Non-Technical	Domain-Specific	Marketing	Empirical	Algorithmic bias	Results
Wexler J., Pushkarna M., Bolukbasi T., Wattenberg M., Viegas F., Wilson J. (2020). The what-if tool: Interactive probing of machine learning models	Technical	Generic	Generic	Empirical	AF	Framework and Metric
Ntoutsi E., Fafalios P., Gadiraju U., Iosifidis V., Nejdl W., Vidal M.-E., Ruggieri S., Turini F., Papadopoulos S., Krasanakis E., Kompatsiaris I., Kinder-Kurlanda K., Wagner C., Karimi F., Fernandez M., Alani H., Berendt B., Kruegel T., Heinze C., Broelemann K., Kasneci G., Tiropanis T., Staab S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey	Non-Technical	Generic	Generic	Empirical	General	Overview
Bellamy R.K.E., Mojsilovic A., Nagar S., Ramamurthy K.N., Richards J., Saha D., Sattigeri P., Singh M., Varshney K.R., Zhang Y., Dey K., Hind M., Hoffman S.C., Houde S., Kannan K., Lohia P., Martino J., Mehta S. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias	Technical	Generic	Generic	Empirical	AF	Framework and Metric
Araujo T., Helberger N., Kruikemeier S., de Vreese C.H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence	Non-Technical	Generic	Generic	Empirical	General	Results
Lee M.K., Jain A., Cha H.J.I.N., Ojha S., Kusbit D. (2019). Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation	Non-Technical	Domain-Specific	Justice	Empirical	AF	Framework and Metric
Ciampaglia G.L. (2018). Fighting fake news: a role for computational social science in the fight against digital misinformation	Non-Technical	Generic	Generic	Conceptual	General	Results
Turner Lee N. (2018). Detecting racial bias in algorithms and machine learning	Non-Technical	Generic	Generic	Empirical	AF	Results
Ahn Y., Lin Y.-R. (2020). Fairsight: Visual analytics for fairness in decision making	Technical	Generic	Generic	Empirical	General	Framework and Metric
Friedler S.A., Scheidegger C., Venkatasubramanian S. (2021). The (Im)possibility of fairness	Technical	Generic	Generic	Conceptual	General	Framework

Paper	Type	Overall Context	Context	Research	Fairness focus	Output
Robert L.P., Pierce C., Marquis L., Kim S., Alahmad R. (2020). Designing fair AI for managing employees in organizations: a review, critique, and design agenda	Non-Technical	Generic	Generic	Conceptual	General	Overview
Galaz V., Centeno M.A., Callahan P. W., Causevic A., Patterson T., Brass I., Baum S., Farber D., Fischer J., Garcia D., McPhearson T., Jimenez D., King B., Larcey P., Levy K. (2021). Artificial intelligence, systemic risks, and sustainability	Technical	Domain-Specific	Sustainability	Empirical	AF	Overview
Wachter S., Mittelstadt B., Russell C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI	Non-Technical	Domain-Specific	Justice	Empirical	AF	Results
Paulus J.K., Kent D.M. (2020). Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities	Technical	Domain-Specific	Medical	Conceptual	AF	Framework
Papakyriakopoulos O., Serrano J.C. M., Hegelich S. (2020). Political communication on social media: A tale of hyperactive users and bias in recommender systems	Technical	Generic	Generic	Empirical	Social Influence Bias	Results
Choudhury P., Starr E., Agarwal R. (2020). Machine learning and human capital complementarities: Experimental evidence on bias mitigation	Technical	Generic	Generic	Empirical	Bias Mitigation	Results
Grgic-Hlaca N., Engel C., Gummadi K.P. (2019). Human decision making with machine advice: An experiment on bailing and jailing	Non-Technical	Domain-Specific	Justice	Empirical	AF	Results
Gu X., Angelov P.P., Soares E.A. (2020). A self-adaptive synthetic over-sampling technique for imbalanced classification	Technical	Generic	Generic	Conceptual	AF	Framework
Helberger N., Huh J., Milne G., Strycharz J., Sundaram H. (2020). Macro and Exogenous Factors in Computational Advertising: Key Issues and New Research Directions	Non-Technical	Domain-Specific	Marketing	Conceptual	AF	Overview
Edizel B., Bonchi F., Hajian S., Panisson A., Tassa T. (2020). FaiRecSys: mitigating algorithmic bias in recommender systems	Technical	Generic	Generic	Empirical	AF	Framework
Shin D. (2021). Embodying algorithms, enactive artificial intelligence and the extended cognition: You can see as much as you know about algorithm	Non-Technical	Generic	Generic	Empirical	General	Results
Yarger L., Cobb Payton F., Neupane B. (2020). Algorithmic equity in the hiring of underrepresented IT job candidates	Non-Technical	Generic	Generic	Empirical	Algorithmic bias	Results

Paper	Type	Overall Context	Context	Research	Fairness focus	Output
Helberger N., Araujo T., de Vreese C.H. (2020). Who is the fairest of them all? Public attitudes and expectations regarding automated decision-making	Non-Technical	Generic	Generic	Empirical	General	Results
Taati B., Zhao S., Ashraf A.B., Asgarian A., Browne M.E., Prkachin K.M., Mihailidis A., Hadjistavropoulos T. (2019). Algorithmic bias in clinical populations - Evaluating and improving facial analysis technology in older adults with dementia	Technical	Domain-Specific	Medical	Empirical	Algorithmic bias	Results
Bandy J. (2021). Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits	Non-Technical	Generic	Generic	Empirical	General	Overview
Bantilan N. (2018). Themis-ml: A Fairness-Aware Machine Learning Interface for End-To-End Discrimination Discovery and Mitigation	Technical	Generic	Generic	Conceptual	General	Framework
Żbikowski K., Antosiuk P. (2021). A machine learning, bias-free approach for predicting business success using Crunchbase data	Technical	Domain-Specific	Economic	Empirical	General	Results
Schramowski P., Turan C., Jentzsch S., Rothkopf C., Kersting K. (2020). The Moral Choice Machine	Technical	Domain-Specific	Ethical	Empirical	General	Framework and Metric
Altman M., Wood A., Vayena E. (2018). A Harm-Reduction Framework for Algorithmic Fairness	Technical	Generic	Generic	Conceptual	AF	Framework
Fletcher R.R., Nakeshimana A., Olubeko O. (2021). Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health	Non-Technical	Domain-Specific	Medical	Empirical	General	Overview
Binns R. (2018). What Can Political Philosophy Teach Us about Algorithmic Fairness?	Non-Technical	Generic	Generic	Conceptual	General	Overview
Bolander T. (2019). What do we loose when machines take the decisions?	Technical	Generic	Generic	Conceptual	General	Overview
Saxena N.A., Huang K., DeFilippis E., Radanovic G., Parkes D.C., Liu Y. (2020). How do fairness definitions fare? Testing public attitudes towards three algorithmic definitions of fairness in loan allocations	Non-Technical	Generic	Generic	Empirical	General	Results
Zehlike M., Hacker P., Wiedemann E. (2020). Matching code and law: achieving algorithmic fairness with optimal transport	Technical	Generic	Generic	Conceptual	AF	Framework
Kozodoi N., Jacob J., Lessmann S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications	Technical	Domain-Specific	Economic	Empirical	General	Results
Checco A., Bracciale L., Loreti P., Pinfield S., Bianchi G. (2021). AI-assisted peer review	Technical	Domain-Specific	Academic	Empirical	Algorithmic bias	Framework

Paper	Type	Overall Context	Context	Research	Fairness focus	Output
Fu R., Huang Y., Singh P.V. (2021). Crowds, lending, machine, and bias	Technical	Domain-Specific	Economic	Empirical	AF	Algorithm and Results
Lee M.S.A., Floridi L. (2021). Algorithmic Fairness in Mortgage Lending: from Absolute Conditions to Relational Trade-offs	Non-Technical	Domain-Specific	Economic	Empirical	AF	Metric
Cheng L., Varshney K.R., Liu H. (2021). Socially responsible AI algorithms: Issues, purposes, and challenges	Non-Technical	Generic	Generic	Conceptual	General	Overview
Gupta M., Parra C.M., Dennehy D. (2021). Questioning Racial and Gender Bias in AI-based Recommendations: Do Espoused National Cultural Values Matter?	Non-Technical	Generic	Generic	Conceptual	General	Overview
Lyu L., Li Y., Nandakumar K., Yu J., Ma X. (2022). How to Democratise and Protect AI: Fair and Differentially Private Decentralised Deep Learning	Technical	Generic	Generic	Empirical	AF	Framework
Valdivia A., Sánchez-Monedero J., Casillas J. (2021). How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness	Technical	Generic	Generic	Empirical	AF	Framework
Grari V., Ruf B., Lamprier S., Detyniecki M. (2020). Achieving Fairness with Decision Trees: An Adversarial Approach	Technical	Generic	Generic	Empirical	AF	Framework and Results
Lyons H., Velloso E., Miller T. (2021). Conceptualising Contestability: Perspectives on Contesting Algorithmic Decisions	Non-Technical	Generic	Generic	Conceptual	General	Overview
Card D., Smith N.A. (2020). On Consequentialism and Fairness	Non-Technical	Generic	Generic	Conceptual	General	Overview
Simon J., Wong P.-H., Rieder G. (2020). Algorithmic bias and the value sensitive design approach	Non-Technical	Generic	Generic	Conceptual	AF	Overview
Fitzsimons J., Al Ali A., Osborne M., Roberts S. (2019). A general framework for fair regression	Technical	Generic	Generic	Empirical	AF	Framework
van Giffen B., Herhausen D., Fahse T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods	Non-Technical	Generic	Generic	Empirical	General	Overview
Madaio M., Egede L., Subramonyam H., Wortman Vaughan J., Wallach H. (2022). Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support	Non-Technical	Generic	Generic	Empirical	General	Overview
Ashokan A., Haas C. (2021). Fairness metrics and bias mitigation strategies for rating predictions	Technical	Generic	Generic	Empirical	AF	Framework and Metric
Miron M., Tolan S., Gómez E., Castillo C. (2021). Evaluating causes of algorithmic bias in juvenile criminal recidivism	Technical	Domain-Specific	Justice	Empirical	Algorithmic bias	Results
Wang Q., Xu Z., Chen Z., Wang Y., Liu S., Qu H. (2021). Visual analysis of discrimination in machine learning	Technical	Generic	Generic	Empirical	AF	Framework

Paper	Type	Overall Context	Context	Research	Fairness focus	Output
Salazar R., Neutatz F., Abedjan Z. (2021). Automated feature engineering for algorithmic fairness	Technical	Generic	Generic	Empirical	AF	Framework
Oneto L., Donini M., Pontil M., Shawe-Taylor J. (2020). Randomized learning and generalization of fair and private classifiers: From PAC-Bayes to stability and differential privacy	Technical	Generic	Generic	Empirical	AF	Framework
Yoon T., Lee J., Lee W. (2020). Joint Transfer of Model Knowledge and Fairness over Domains Using Wasserstein Distance	Technical	Generic	Generic	Conceptual	AF	Algorithm and Results
Hamon R., Junklewitz H., Sanchez I., Malgieri G., De Hert P. (2022). Bridging the Gap between AI and Explainability in the GDPR: Towards Trustworthiness-by-Design in Automated Decision-Making	Non-Technical	Generic	Generic	Conceptual	General	Overview
Zajko M. (2021). Conservative AI and social inequality: conceptualizing alternatives to bias through social theory	Non-Technical	Domain-Specific	Justice	Empirical	AF	Overview
de Sousa I.P., Vellasco M.M.B.R., da Silva E.C. (2021). Explainable artificial intelligence for bias detection in covid ct-scan classifiers	Technical	Domain-Specific	Medical	Empirical	AF	Results
Georgopoulos M., Oldfield J., Nicolaou M.A., Panagakis Y., Pantic M. (2021). Mitigating Demographic Bias in Facial Datasets with Style-Based Multi-attribute Transfer	Technical	Generic	Generic	Empirical	Dataset Bias	Framework
Dolata M., Feuerriegel S., Schwabe G. (2022). A sociotechnical view of algorithmic fairness	Technical	Generic	Generic	Empirical	AF	Overview
Sartori L., Theodorou A. (2022). A sociotechnical perspective for the future of AI: narratives, inequalities, and human control	Technical	Generic	Generic	Empirical	AF	Overview
Makhlouf K., Zhioua S., Palamidessi C. (2021). Machine learning fairness notions: Bridging the gap with real-world applications	Non-Technical	Generic	Generic	Conceptual	General	Framework and Metric
Fernando M.-P., Cèsar F., David N., José H.-O. (2021). Missing the missing values: The ugly duckling of fairness in machine learning	Technical	Generic	Generic	Empirical	AF	Results
Varley M., Belle V. (2021). Fairness in machine learning with tractable models	Technical	Generic	Generic	Empirical	AF	Framework
Zhang T., Zhu T., Gao K., Zhou W., Yu P.S. (2021). Balancing Learning Model Privacy, Fairness, and Accuracy With Early Stopping Criteria	Technical	Generic	Generic	Empirical	AF	Framework
Morse L., Teodorescu M.H.M., Awwad Y., Kane G.C. (2021). Do the Ends Justify the Means? Variation in the Distributive and Procedural Fairness of Machine Learning Algorithms	Technical	Generic	Generic	Conceptual	AF	Framework

Paper	Type	Overall Context	Context	Research	Fairness focus	Output
Kehrenberg T., Chen Z., Quadrianto N. (2020). Tuning Fairness by Balancing Target Labels	Technical	Generic	Generic	Empirical	AF	Framework
Zhang K., Khosravi B., Vahdati S., Faghani S., Nugen F., Rassoulinejad-Mousavi S.M., Moassefi M., Jagtap J.M.M., Singh Y., Rouzrokh P., Erickson B.J. (2022). Mitigating Bias in Radiology Machine Learning: 2. Model Development	Technical	Domain-Specific	Medical	Empirical	AF	Framework
Lewis A., Stoyanovich J. (2022). Teaching Responsible Data Science: Charting New Pedagogical Territory	Non-Technical	Domain-Specific	Education	Conceptual	AF	Overview
von Zahn M., Feuerriegel S., Kuehl N. (2022). The Cost of Fairness in AI: Evidence from E-Commerce	Technical	Domain-Specific	Economic	Empirical	AF	Algorithm and Results
Akter S., Dwivedi Y.K., Sajib S., Biswas K., Bandara R.J., Michael K. (2022). Algorithmic bias in machine learning-based marketing models	Technical	Domain-Specific	Marketing	Empirical	Algorithmic bias	Framework
Bærøe K., Gundersen T., Henden E., Rommetveit K. (2022). Can medical algorithms be fair? Three ethical quandaries and one dilemma	Non-Technical	Domain-Specific	Medical	Conceptual	AF	Overview
Yee K., Tantipongpipat U., Mishra S. (2021). Image Cropping on Twitter: Fairness Metrics, their Limitations, and the Importance of Representation, Design, and Agency	Technical	Generic	Generic	Empirical	AF	Metric
Das S., Donini M., Gelman J., Haas K., Hardt M., Katzman J., Kenthapadi K., Larroy P., Yilmaz P., Zafar M.B. (2021). Fairness Measures for Machine Learning in Finance	Technical	Domain-Specific	Economic	Empirical	AF	Metric
Alelyani S. (2021). Detection and evaluation of machine learning bias	Technical	Generic	Generic	Empirical	Detection	Framework
Li J., Moskovitch Y., Jagadish H.V. (2021). Denouncer: Detection of unfairness in classifiers	Technical	Generic	Generic	Empirical	AF	Framework
Oneto L. (2020). Learning fair models and representations	Technical	Generic	Generic	Conceptual	AF	Framework
Wang M., Zhang Y., Deng W. (2022). Meta Balanced Network for Fair Face Recognition	Technical	Generic	Generic	Empirical	AF	Algorithm and Results
Bogina V., Hartman A., Kuflik T., Shulner-Tal A. (2022). Educating Software and AI Stakeholders About Algorithmic Fairness, Accountability, Transparency and Ethics	Non-Technical	Domain-Specific	Education	Empirical	AF	Overview
Fu R., Aseri M., Singh P.V., Srinivasan K. (2022). "Un"Fair Machine Learning Algorithms	Technical	Generic	Generic	Empirical	AF	Results
Franco D., Navarin N., Donini M., Anguita D., Oneto L. (2022). Deep fair models for complex data: Graphs labeling and explainable face recognition	Technical	Generic	Generic	Empirical	AF	Framework and Metric

Paper	Type	Overall Context	Context	Research	Fairness focus	Output
Giovanola B., Tiribelli S. (2022). Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms	Non-Technical	Domain-Specific	Medical	Empirical	General	Overview
Soremekun E., Udeshi S.S., Chattopadhyay S. (2022). ASTRAEA: Grammar-based Fairness Testing	Technical	Generic	Generic	Empirical	General	Framework
Mengesha Z., Heldreth C., Lahav M., Sublewski J., Tuennerman E. (2021). "I don't Think These Devices are Very Culturally Sensitive."—Impact of Automated Speech Recognition Errors on African Americans	Non-Technical	Generic	Generic	Empirical	AF	Framework
Criado N., Ferrer X., Such J.M. (2021). Attesting digital discrimination using norms	Non-Technical	Generic	Generic	Empirical	AF	Overview
Yang T., Yao R., Yin Q., Tian Q., Wu O. (2021). Mitigating sentimental bias via a polar attention mechanism	Technical	Generic	Generic	Empirical	Mitigation	Algorithm and Results
Ahmed Z., Vidgen B., Hale S.A. (2022). Tackling racial bias in automated online hate detection: Towards fair and accurate detection of hateful users with geometric deep learning	Technical	Generic	Generic	Empirical	AF	Framework
Chen M.-Y., Chiang H.-S., Huang W.-K. (2022). Efficient Generative Adversarial Networks for Imbalanced Traffic Collision Datasets	Technical	Generic	Generic	Empirical	AF	Results
Faghani S., Khosravi B., Zhang K., Moassefi M., Jagtap J.M., Nugen F., Vahdati S., Kuanar S.P., Rassoulinejad-Mousavi S.M., Singh Y., Vera Garcia D.V., Rouzrokh P., Erickson B.J. (2022). Mitigating Bias in Radiology Machine Learning: 3. Performance Metrics	Technical	Domain-Specific	Medical	Empirical	AF	Metric
Anahideh H., Asudeh A., Thirumuruganathan S. (2022). Fair active learning	Technical	Generic	Generic	Empirical	AF	Framework and Metric
Nakao Y., Stumpf S., Ahmed S., Naseer A., Strappelli L. (2022). Toward Involving End-users in Interactive Human-in-the-loop AI Fairness	Non-Technical	Domain-Specific	Economic	Empirical	General	Results
Hazirbas C., Bitton J., Dolhansky B., Pan J., Gordo A., Ferrer C.C. (2022). Towards Measuring Fairness in AI: The Casual Conversations Dataset	Technical	Generic	Generic	Empirical	AF	Overview
Aler Tubella A., Barsotti F., Koçer R. G., Mendez J.A. (2022). Ethical implications of fairness interventions: what might be hidden behind engineering choices?	Non-Technical	Domain-Specific	Ethical	Conceptual	AF	Framework and Metric

Paper	Type	Overall Context	Context	Research	Fairness focus	Output
Jain N., Olmo A., Sengupta S., Manikonda L., Kambhampati S. (2022). Imperfect ImaGANation: Implications of GANs exacerbating biases on facial data augmentation and snapchat face lenses	Technical	Generic	Generic	Empirical	AF	Results
Petrović A., Nikolić M., Radovanović S., Delibašić B., Jovanović M. (2022). FAIR: Fair adversarial instance re-weighting	Technical	Generic	Generic	Empirical	AF	Framework
Fabris A., Messina S., Silvello G., Susto G.A. (2022). Algorithmic fairness datasets: the story so far	Non-Technical	Generic	Generic	Empirical	AF	Overview
Pessach D., Shmueli E. (2021). Improving fairness of artificial intelligence algorithms in Privileged-Group Selection Bias data settings	Technical	Generic	Generic	Empirical	Algorithmic Bias	Framework
Radovanović S., Petrović A., Delibašić B., Suknović M. (2021). A fair classifier chain for multi-label bank marketing strategy classification	Technical	Domain-Specific	Economic	Empirical	AF	Algorithm and Results
Duan C.J., Gaurav A. (2021). Exposing model bias in machine learning revisiting the boy who cried wolf in the context of phishing detection	Technical	Generic	Generic	Empirical	AF	Metric
Plečko D., Meinshausen N. (2020). Fair data adaptation with quantile preservation	Technical	Generic	Generic	Empirical	AF	Framework
Castelnovo A., Cosentini A., Malandri L., Mercorio F., Mezzanzanica M. (2022). FFTree: A flexible tree to handle multiple fairness criteria	Technical	Generic	Generic	Conceptual	AF	Framework and Metric
Sokol K., Santos-Rodriguez R., Flach P. (2022). FAT Forensics: A Python toolbox for algorithmic fairness, accountability and transparency[Formula presented]	Technical	Generic	Generic	Conceptual	AF	Framework
Kim K., Ohn I., Kim S., Kim Y. (2022). SLIDE: A surrogate fairness constraint to ensure fairness consistency	Technical	Generic	Generic	Empirical	AF	Framework
Kim J.-Y., Cho S.-B. (2022). An information theoretic approach to reducing algorithmic bias for machine learning	Technical	Generic	Generic	Empirical	AF	Metric
Folorunso S., Ogundepo E., Basajja M., Awotunde J., Kawu A., Oladipo F., Abdullahi I. (2022). FAIR Machine Learning Model Pipeline Implementation of COVID-19 Data	Technical	Domain-Specific	Medical	Empirical	AF	Framework
Wang A., Liu A., Zhang R., Kleiman A., Kim L., Zhao D., Shirai I., Narayanan A., Russakovsky O. (2022). REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets	Technical	Generic	Generic	Empirical	General	Framework

Paper	Type	Overall Context	Context	Research	Fairness focus	Output
Risser L., Sanz A.G., Vincenot Q., Loubes J.-M. (2022). Tackling Algorithmic Bias in Neural-Network Classifiers using Wasserstein-2 Regularization	Technical	Generic	Generic	Empirical	Algorithmic Bias	Algorithm and Results
Liu S., Vicente L.N. (2022). Accuracy and fairness trade-offs in machine learning: a stochastic multi-objective approach	Technical	Generic	Generic	Empirical	AF	Framework
Blanzeisky W., Cunningham P. (2022). Using Pareto simulated annealing to address algorithmic bias in machine learning	Technical	Generic	Generic	Empirical	Algorithmic Bias	Results
Mosteiro P., Kuiper J., Masthoff J., Scheepers F., Spruit M. (2022). Bias Discovery in Machine Learning Models for Mental Health	Technical	Domain-Specific	Medical	Empirical	Algorithmic Bias	Metric
Lee J., Bu Y., Sattigeri P., Panda R., Wornell G.W., Karlinsky L., Feris R. S. (2022). A Maximal Correlation Framework for Fair Machine Learning	Technical	Generic	Generic	Empirical	AF	Results
Li S., Yu J., Du X., Lu Y., Qiu R. (2022). Fair Outlier Detection Based on Adversarial Representation Learning	Technical	Generic	Generic	Empirical	General	Framework
Li Y., Huang H., Geng Q., Guo X., Yuan Y. (2022). Fairness Measures of Machine Learning Models in Judicial Penalty Prediction	Technical	Domain-Specific	Justice	Empirical	AF	Metric
Zhang Q., Liu J., Zhang Z., Wen J., Mao B., Yao X. (2022). Mitigating Unfairness via Evolutionary Multi-objective Ensemble Learning	Technical	Generic	Generic	Empirical	AF	Framework
Li C., Xing W., Leite W. (2022). Using fair AI to predict students' math learning outcomes in an online platform	Technical	Domain-Specific	Education	Empirical	AF	Algorithm and Results
Caton S., Malisetty S., Haas C. (2022). Impact of Imputation Strategies on Fairness in Machine Learning	Non-Technical	Generic	Generic	Empirical	General	Results
Sha L., Rakovic M., Das A., Gasevic D., Chen G. (2022). Leveraging Class Balancing Techniques to Alleviate Algorithmic Bias for Predictive Tasks in Education	Technical	Domain-Specific	Education	Empirical	Algorithmic Bias	Framework
Strobel M., Shokri R. (2022). Data Privacy and Trustworthy Machine Learning	Non-Technical	Generic	Generic	Conceptual	General	Overview
Ferry J., Aïvodji U., Gambs S., Huguet M.-J., Siala M. (2022). Improving fairness generalization through a sample-robust optimization method	Technical	Generic	Generic	Empirical	Statistical Fairness	Framework
Curto G., Jojoa Acosta M.F., Comim F., Garcia-Zapirain B. (2022). Are AI systems biased against the poor? A machine learning analysis using Word2Vec and GloVe embeddings	Technical	Generic	Generic	Empirical	AF	Results

Paper	Type	Overall Context	Context	Research	Fairness focus	Output
Nakao Y., Strappelli L., Stumpf S., Naseer A., Regoli D., Gamba G.D. (2022). Towards Responsible AI: A Design Space Exploration of Human-Centered Artificial Intelligence User Interfaces to Investigate Fairness	Non-Technical	Domain-Specific	Economic	Empirical	General	Framework
Waller R.R., Waller R.L. (2022). Assembled Bias: Beyond Transparent Algorithmic Bias	Non-Technical	Generic	Generic	Conceptual	AF	Overview
Weinberg L. (2022). Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches	Non-Technical	Generic	Generic	Empirical	General	Overview
Liu S., Sun S., Zhao J. (2022). Fair Transfer Learning with Factor Variational Auto-Encoder	Technical	Generic	Generic	Empirical	AF	Framework
Kairouz P., Liao J., Huang C., Vyas M., Welfert M., Sankar L. (2022). Generating Fair Universal Representations Using Adversarial Models	Technical	Generic	Generic	Conceptual	AF	Framework and Metric
Andreeva O., Almeida M., Ding W., Crouter S.E., Chen P. (2022). Maximizing Fairness in Deep Neural Networks via Mode Connectivity	Technical	Generic	Generic	Conceptual	AF	Framework
Saraswat A., Pal M., Pokhriyal S., Abhishek K. (2022). Towards fair machine learning using combinatorial methods	Non-Technical	Generic	Generic	Conceptual	AF	Overview
Zhao H., Gordon G.J. (2022). Inherent Tradeoffs in Learning Fair Representations	Technical	Generic	Generic	Empirical	AF	Algorithm and Results
Scantamburlo T. (2021). Non-empirical problems in fair machine learning	Non-Technical	Generic	Generic	Conceptual	AF	Overview
Nguyen D., Gupta S., Rana S., Shilton A., Venkatesh S. (2021). Fairness improvement for black-box classifiers with Gaussian process	Technical	Generic	Generic	Empirical	AF	Framework and Metric
Vermeer N., Boer A., Winkels R. (2021). Survivorship bias mitigation in a recidivism prediction tool	Technical	Generic	Generic	Empirical	AF	Framework
Dehouche N. (2021). Implicit Stereotypes in Pre-Trained Classifiers	Technical	Generic	Generic	Empirical	AF	Results
Kanamori K., Arimura H. (2021). Fairness-aware decision tree editing based on mixedinteger linear optimization	Technical	Generic	Generic	Empirical	AF	Framework and Metric
Popa A. (2021). Fairness Embedded Adaptive Recommender System: A Conceptual Framework	Technical	Generic	Generic	Empirical	AF	Framework and Metric
Nieto, N., Larrazabal, A., Peterson, V., Milone, D.H., Ferrante, E. (2021). On the relationship between research parasites and fairness in machine learning: Challenges and opportunities	Non-Technical	Generic	Generic	Conceptual	AF	Overview

Sikt Approval and Information Letter

.1 Sikt Approval

Sikt

Notification form / Masteroppgave om algoritmisk rettferdighet / Assessment

Assessment of processing of personal data

Reference number	Assessment type	Date
694968	Automatic ?	01.03.2023

Project title
Masteroppgave om algoritmisk rettferdighet

Data controller (institution responsible for the project)
Norges teknisk-naturvitenskapelige universitet / Fakultet for informasjonsteknologi og elektroteknikk (IE) / Institutt for datateknologi og informatikk

Project leader
Ilias O. Pappas

Student
Fredrik Wilhelm Butler Wang

Project period
06.03.2023 - 12.06.2023

Categories of personal data
General

Legal basis
Consent (General Data Protection Regulation art. 6 nr. 1 a)

The processing of personal data is lawful, so long as it is carried out as stated in the notification form. The legal basis is valid until 12.06.2023.

Notification Form

Basis for automatic assessment
The notification form has received an automatic assessment. This means that the assessment has been automatically generated based on the information registered in the notification form. Only processing of personal data with low risk for data subjects receive an automatic assessment. Key criteria are:

- Data subjects are over the age of 15
- Processing does not include special categories of personal data;
  - Racial or ethnic origin
  - Political, religious or philosophical beliefs
  - Trade union membership
  - Genetic data
  - Biometric data to uniquely identify an individual
  - Health data
  - Sex life or sexual orientation
- Processing does not include personal data about criminal convictions and offences
- Personal data shall not be processed outside the EU/EEA, and no one located outside the EU/EEA shall have access to the personal data
- Data subjects will receive information in advance about the processing of their personal data.

Information provided to data subjects (samples) must include

- The identity and contact details of the data controller
- Contact details of the data protection officer (if relevant)
- The purpose for processing personal data
- The scientific purpose of the project
- The legal basis for processing personal data
- What type of personal data will be processed and how it will be collected, or from where it will be obtained
- Who will have access to the personal data (categories of recipients)

- How long the personal data will be processed
- The right to withdraw consent and other rights

We recommend using our template for the information letter.

Information security
You must process the personal data in accordance with the storage guide and information security guidelines of the data controller. The institution is responsible for ensuring that the conditions of Article 5(1)(d) accuracy and 5(1)(f) integrity and confidentiality, as well as Article 32 security, are met.

.2 Information Letter

Are you interested in taking part in the research project

"Implementation of algorithmic fairness"?

Purpose of the project
You are invited to participate in a research project where the main purpose is to *figure out how companies and organisations implement algorithmic fairness*.

The main goal will be to figure out how algorithmic fairness is implemented by companies and organisations.
The project's objectives include:

- What approach is taken to implement algorithmic fairness?
- How was this approach chosen?
- Is there a form of framework for algorithmic fairness that is implemented?
- Are there further steps/measures that could be taken?

This is project is a part of a master's thesis at Norwegian University of Science and Technology via the Faculty of Information Technology and Electrical Engineering / Department of Computer Science

Which institution is responsible for the research project?
Norwegian University of Science and Technology Faculty of Information Technology and Electrical Engineering / Department of Computer Science is responsible for the project (data controller).

Why are you being asked to participate?
We wish to interview people who work with AI systems and automated decision-making systems where fairness is implemented. We wish to interview between 10-20 people.

What does participation involve for you?
If you chose to take part in the project, this will involve that you participate in an online interview. It will take approximately 30-60 minutes. The interview includes questions about algorithmic fairness. It will also consist of some general questions such as your background/role. There will be a sound and video recording of the interview that is only for internal use and that later will be deleted.

Participation is voluntary
Participation in the project is voluntary. If you chose to participate, you can withdraw your consent at any time without giving a reason. All information about you will then be made anonymous. There will be no negative consequences for you if you chose not to participate or later decide to withdraw.

Your personal privacy – how we will store and use your personal data
We will only use your personal data for the purpose(s) specified here and we will process your personal data in accordance with data protection legislation (the GDPR).
Only researchers associated with the project will have access to the data. Your name and contact details will be replaced with a code. The list of names, contact details and respective codes will be stored separately from the rest of the collected data. In the publication, your role and the size and type of company (but not the company name) will be published. All other information will be anonymized.

What will happen to your personal data at the end of the research project?
The planned end date of the project is 12.06.2023. After the end of the project the data material with your personal data, including any digital recordings, will be anonymized, so that you are not recognisable. This is done for verification, follow-up studies and potential future research.

Your rights
So long as you can be identified in the collected data, you have the right to:

- access the personal data that is being processed about you
- request that your personal data is deleted
- request that incorrect personal data about you is corrected/rectified
- receive a copy of your personal data (data portability), and
- send a complaint to the Norwegian Data Protection Authority regarding the processing of your personal data

What gives us the right to process your personal data?
We will process your personal data based on your consent.

Based on an agreement with Norwegian University of Science and Technology *Faculty of Information Technology and Electrical Engineering / Department of Computer Science,* The Data Protection Services of Sikt – Norwegian Agency for Shared Services in Education and Research has assessed that the processing of personal data in this project meets requirements in data protection legislation.

Where can I find out more?
If you have questions about the project, or want to exercise your rights, contact:

- Norwegian University of Science and Technology via
  Project Leader / Supervisor: Ilias O. Pappas (ilpappas@ntnu.no , 73594427)
  Student: Fredrik Wilhelm Butler Wang (fredrikbw@outlook.com , 48355990)
- Our Data Protection Officer: Thomas Helgesen (Thomas.helgesen@ntnu.no)

If you have questions about how data protection has been assessed in this project by Sikt, contact:

- email: (personverntjenester@sikt.no) or by telephone: +47 73 98 40 40.

Yours sincerely,

Ilias O. Pappas
Project Leader
(Researcher/supervisor)

Fredrik Wilhelm Butler Wang
Student

Consent form

I have received and understood information about the project Implementation of algorithmic fairness and have been given the opportunity to ask questions. I give consent:

- □ To participate in an interview
- □ To allow for audio recording of the interviews
- □ To allow for video recording of the interviews

I give consent for my personal data to be processed until the end of the project.

(Signed by participant, date)

Interview Guide

ENGLISH: Interview Guide

Background:

- What is your background?
- What is your role at this company/organization?
- Can you briefly explain what type of algorithmic decision making you/your team is utilizing?
- How much experience do you have with AF / fairness?

General:

- What does fairness in general mean to you?
- What does algorithmic fairness mean to you?
- According to you, when is algorithmic fairness achieved?

<Short explanation of fairness>

Any case where AI/ML systems perform differently for different groups in ways that may be considered undesirable.

<Explain what is meant by "framework">

Definition of fairness framework:

A fairness framework refers to a set of tools, standards and conventions that provide a foundation for tackling fairness in a technical solution. In the context of AF, it includes methods for mitigating, detecting, and handling AF. Software toolkits and checklists are common tools.

Frameworks for AF:

- Do you apply any framework(s) for achieving algorithmic fairness?

IF THEY DO:

  - What does it mean that you apply this framework?
    - Are you following any rules or steps?
  - How did you decide on this framework?
  - Is the framework implementing algorithmic fairness throughout the pipeline of your solution?
  - What challenges do you have with this framework?
  - Is there anything with the framework you are not implementing? If so, what are the reasons for that?
  - Is there anything else you would like to tell us about this framework or your process for implementing AF?

- **IF THEY DON'T:**
  - What do you do to detect algorithmic unfairness?
  - What approaches do you take to mitigate algorithmic unfairness?
  - How did you decide on these approaches?
  - Are you implementing algorithmic fairness throughout the pipeline of your solution?
  - Is there something you are not doing that you could do to mitigate algorithmic unfairness?
  - What are the challenges you have with detecting or mitigating AF?
  - Do you think that you could benefit from applying a framework for AF?

NORSK: Intervju guide

Bakgrunn:

- Hva er bakgrunnen din?
- Hva er din rolle i dette selskapet/organisasjonen?
- Kan du kort forklare hva slags algoritmisk beslutningstagning som du / ditt team benytter seg av?
- Hvor mye erfaring har du med algoritmisk rettferdighet / rettferdighet

General:

- Helt generelt, hva betyr rettferdighet for deg?
- Hva betyr algoritmisk rettferdighet for deg?
- Ifølge deg selv, når er algoritmisk rettferdighet oppnådd?

<Kort forklaring av rettferdighet>

Alle tilfeller der AI/ML-systemer fungerer ulikt for ulike grupper på måter som kan anses som uønskede.

<Forklar hva som menes med «rammeverk»>

Definisjon av rettferdighetsrammeverk:

Et rettferdighetsrammeverk refererer til et sett med verktøy, standarder og konvensjoner som gir et grunnlag for å håndtere rettferdighet i en teknisk løsning. I sammenheng med AF inkluderer det metoder for å dempe, oppdage og håndtere AF. Programvareverktøysett og sjekklister er vanlige verktøy.

Rammeverk for AF:

- Bruker du noen rammeverk for å oppnå algoritmisk rettferdighet?

  HVIS DE GJØR:

    - Hva betyr det at du bruker dette rammeverket?
        - Følger du noen regler eller trinn?
    - Hvordan bestemte du deg for dette rammeverket?
    - Implementerer rammeverket algoritmisk rettferdighet gjennom hele pipelinen til løsningen din?
    - Hvilke utfordringer har du med dette rammeverket?
    - Er det noe med rammeverket du ikke implementerer? I så fall, hva er årsakene til det?
    - Er det noe annet du vil fortelle oss om dette rammeverket eller prosessen din for implementering av AF?
- **HVIS DE IKKE GJØR:**
    - Hva gjør du for å oppdage algoritmisk urettferdighet?
    - Hvilke tilnærminger bruker du for å redusere algoritmisk urettferdighet?
    - Hvordan bestemte du deg for disse tilnærmingene?
    - Implementerer du algoritmisk rettferdighet gjennom hele løsningen?
    - Er det noe du ikke gjør som du kan gjøre for å redusere algoritmisk urettferdighet?
    - Hva er utfordringene du har med å oppdage eller dempe AF?
    - Tror du at du kan ha nytte av å bruke et rammeverk for AF?

NTNU

Norwegian University of
Science and Technology