Aalto-yliopisto
Aalto-universitetet
Aalto University

Bachelor’s Programme in Information and Service Management

Preventing unfair outcomes caused by algorithmic bias

Socioethical view into the issues of algorithmic bias in artificial intelligence

Lari Kauranen

Bachelor’s thesis
2024

Aalto-yliopisto
Aalto-universitetet
Aalto University

Author Lari Kauranen

Title of thesis Preventing unfair outcomes caused by algorithmic bias

Programme Bachelor's degree

Major Information and Service Management

Thesis supervisor Johanna Bragge

Thesis advisor(s) Johanna Bragge

Date 04.11.2024 **Number of pages** 28 **Language** English

Abstract

Algorithms and AI increasingly impact people, organizations, and nations. A lot of the past research has focused on the positive effects of these technologies but recently there has also been an increasing amount of attention directed towards the possible negative effects and the scope of the issue. Negative effects may potentially include things like increases in biases, destructive effects, reduction of human agency, spread of misinformation and propaganda trough deepfakes, privacy issues and transparency issues.

This thesis inspects the issues of unfair outcomes caused by algorithmic bias through a literature review aiming to bring together research from various fields due to the complex and multidisciplinary nature of the issue. The 3-D Dependable AI Framework is utilized as a tool to showcase possible ways of preventing unfair outcomes.

Different types of biases, broadly categorized into cognitive and technical biases, are identified and presented as potential causes of unfair outcomes. The autonomy of AI systems and shortcomings in legal and regulatory frameworks are also examined and highlighted as potential causes of unfair outcomes, warranting further attention.

Possible ways of preventing unfair outcomes are examined and introduced with the help of the 3-D Dependable AI Framework which divides the life cycle of AI into 3 main stages: design, develop and deploy. Each of these stages hold important steps in developing and maintaining ethical AI.

The limitations of the research are discussed, along with implications for practice, emphasizing the need for action and further examination from both governmental and organizational perspectives. Directions and recommendations for future research are also presented.

Keywords Artificial Intelligence, Bias, Algorithm, Algorithmic bias, Ethical AI

Aalto-yliopisto
Aalto-universitetet
Aalto University

Tekijä Lari Kauranen	
Työn nimi Algoritmisten harhojen aiheuttamien haittavaikutuksien ehkäiseminen.	
Koulutusohjelma Kandidaatin tutkinto	
Pääaine Tieto- ja palvelujohtaminen	
Vastuuopettaja/valvoja Johanna Bragge	
Työn ohjaaja(t) Johanna Bragge	
Päivämäärä 04.11.2024 **Sivumäärä** 28	**Kieli** Englanti

Tiivistelmä

Tekoäly ja algoritmit vaikuttavat lisääntyvissä määrin ihmisten, organisaatioiden ja jopa valtioiden toimintaan. Suurin osa aikaisemmasta tutkimuksesta on keskittynyt näiden järjestelmien positiivisiin vaikutuksiin, mutta lähiaikoina myös tekoälyn mahdollisten negatiivisten vaikutusten sekä ongelmien laajuuden tarkastelu ovat lisääntyneet. Mahdolliset negatiiviset vaikutukset voivat sisältää esimerkiksi algoritmisten harhojen lisääntyminen, tuhoisia vaikutuksia, ihmisen toimijuuden vähentyminen, väärän tiedon ja propagandan leviäminen deepfake teknologian avulla, yksityisyysoikeuksien heikkeneminen ja toiminnallisen läpinäkyvyyden heikkeneminen.

Tämä kandidaatintyö tarkastelee algoritmisten harhojen aiheuttamien epäoikeudenmukaisuuksien ehkäisemistä kirjallisuuskatsauksen avulla. Työ pyrkii yhdistämään tutkimusta monelta eri alalta, aiheen monimutkaisuuden ja monitieteisyyden takia. 3-D Dependable AI Framework viitekehystä hyödynnetään työkaluna ja sen mahdollisuuksia ehkäistä haittavaikutuksia tuodaan esille löydöksien avulla.

Työ tunnistaa karkeasti jaoteltuna kahden tyyppisiä vinoumia, kognitiivisia sekä teknisiä vinoumia, joiden osoitetaan aiheuttavan algoritmisia vinoumia ja negatiivisia haittavaikutuksia. Tekoälyjärjestelmien autonomia sekä lainsäädännön ja sääntelykehysten puutteet tunnistetaan myös mahdollisiksi syiksi tekoälyn aiheuttamille haittavaikutuksille.

Mahdollisia tapoja ehkäistä tekoälyn aiheuttamia haittavaikutuksia tarkastellaan ja esitellään 3-D Dependable AI Framework viitekehystä avuksi käyttäen. Viitekehys jakaa tekoälyn kehittämis- sekä tarkasteluprosessin kolmeen päävaiheeseen: suunnittelu, kehittäminen ja käyttöönotto, jotka jokainen sisältävät tärkeitä vaiheita eettisen tekoälyn kehittämistä ja ylläpitämistä varten.

Lopuksi tutkimuksen rajoitteista keskustellaan, sekä vaikutuksista käytännölle korostaen tarvetta toiminnalle ja tarkastelulle valtioiden sekä organisaatioiden näkökulmasta. Myös suuntia sekä suosituksia tulevaisuuden tutkimukselle esitetään.

Avainsanat Tekoäly, algoritmi, vinouma, algoritminen vinouma, eettinen tekoäly

Table of contents

1 Introduction ... 6
1.1 Research objectives and research questions ... 6
1.2 Scope of research ... 6
1.3 Methodology ... 7
1.4 Structure of the research ... 8
2 Key concepts for unfair outcomes caused by algorithmic bias ... 10
2.1 Algorithmic bias ... 10
2.2 Defining unfairness ... 10
2.3 3-D Dependable AI Framework ... 11
3 Unfair outcomes caused by algorithmic bias ... 13
3.1 Differing sources of bias ... 13
3.1.1 Cognitive biases ... 13
3.1.2 Technical biases ... 14
3.2 Autonomy of the system ... 14
3.3 Legal and regulatory shortcomings ... 15
3.3.1 Self-governance and lack of standardized practices ... 16
3.3.2 Speed of progress ... 16
4 Preventing unfair outcomes caused by algorithmic bias ... 17
4.1 Design phase ... 17
4.2 Develop phase ... 18
4.2.1 Data selection and preparation ... 19
4.2.2 Model training and evaluation ... 19
4.3 Deploy phase ... 20
4.3.1 Algorithm accessibility ... 21
4.3.2 Intended use and fairness monitoring ... 21
5 Conclusions ... 24
5.1 Limitations ... 24
5.2 Implications for practice ... 24
5.3 Future research ... 25
References ... 26

1 Introduction

Algorithms and artificial intelligence have had an increasing effect on the daily lives of people, companies, organizations and even countries (Buolamwini, 2018; Mikalef et al., 2022; Rinta-Kahila et al., 2022). Interest regarding their use cases has skyrocketed and it seems that they are coming to stay. With the increasing usage of algorithms their influence on human life also increases. A lot of past research has focused on how these technologies improve people's lives (Hagendorff, 2020) but in the latest years there has also been an increasing amount of attention directed towards the possible negative effects and the scope of the issue (Bellamy et al., 2019; Mikalef et al., 2022).

These negative effects may potentially include things like increases in biases and unfair outcomes, destructive effects for different kinds of stakeholders, reduction of human agency, spread of misinformation and propaganda through deepfakes (Hagendorff, 2020), privacy issues and transparency issues (Mikalef et al., 2022; Rinta-Kahila et al., 2022; Wirtz et al., 2020) just to name a few. These issues are complex and multidimensional. Sources for different kinds of biases can be divided into categories that include at least human bias, machine bias, systemic bias, societal bias, historical bias, sampling bias and observation bias (Zajko, 2021).

1.1 Research objectives and research questions

The objective of this thesis is to scope out the causes of algorithmic bias and to consolidate the existing literature regarding preventing unfair outcomes caused by algorithmic bias when an algorithm is used in artificial intelligence that affects people's lives.

The specific research questions of this thesis are the following:

- How do biases in algorithms cause unfair outcomes?
- How to prevent unfair outcomes caused by algorithmic bias?

As the issues are multidisciplinary, this thesis also tries to tie together the findings from multiple fields from the more technological fields of, for example, information systems to the more sociological fields and to emphasize the need of multidisciplinary co-operation in future research.

1.2 Scope of research

Mittelstadt et al. (2016) introduces six types of ethical concerns raised by algorithms. The scope of research of this thesis is narrowed down from the multiple ethical concerns raised by algorithms to just the **unfair outcomes** (Mittelstadt et al., 2016).

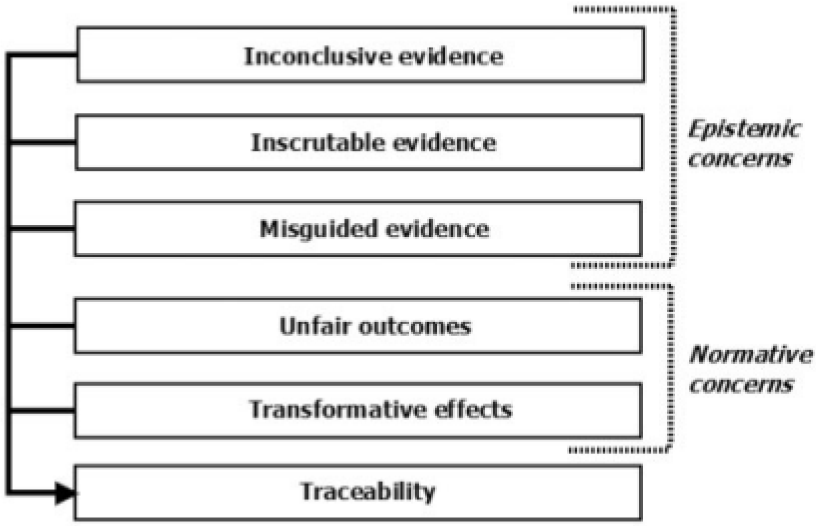

Figure 1. Six types of ethical concerns raised by algorithms (Mittelstadt et al., 2016)

The research is also only going to focus on algorithms that are used in artificial intelligence that affects the lives of people in some ways with the focus on the more major impacts that may include for example upholding racial bias that causes unequal access to healthcare (Obermeyer et al., 2019) or unfair treatment at stages of the criminal process (Grunwald, 2023).

As the length of a Bachelor's thesis is limited and the more technical and mathematical side of algorithms and artificial intelligence is incredibly complex this thesis does not delve deep into the technical or mathematical aspects and theories behind algorithms and artificial intelligence.

1.3 Methodology

The method used in this thesis is a literature review that aims to answer the research questions and bring together research from multiple fields to better understand the complex issues regarding algorithmic bias. The paper utilizes a "socioethical" view of the issue and seeks to integrate diverse findings from past research.

With the multidisciplinarity of the issue, sources were found from various fields including for example criminology, psychology, sociology, information systems, law, philosophy, big data, machine learning, computer science and data mining.

To ensure the credibility of the referenced papers they had to fulfill at least one of the following conditions:

- JUFO portal score of at least 1
- Field-Weighted Citation Impact score of at least 1 in Scopus
- 30 citations or more

With the rapid development of algorithms and artificial intelligence (Buolamwini, 2018; Mittelstadt et al., 2016), there was an effort to also pick out articles that have been published in the recent years so as not to fall behind the newest leaps of progress and technological development.

Research was searched mainly from Scopus with keywords such as “bias*”, “inequality*”, “fairness”, “unfair”, “artificial intelligence”, “algorithm”, “algorithmic fairness”, “ethics”, “discrimination”, “sociology”, “machine learning”, “debias*” and “AI”. Scopus AI was also utilized in gaining new insights and standpoints regarding the topic of algorithmic bias. Backward and forward reference searching were also significant factors in finding key publications.

The research is also going to utilize the 3-D Dependable AI Framework (Ukanwa, 2024) throughout the whole thesis and tie together different findings within the framework and showcase how it could be utilized in preventing unfair outcomes caused by bias.

1.4 Structure of the research

This thesis consists of 4 sections after the introduction. Section 2 introduces some theoretical backgrounds regarding algorithmic bias and elaborates on the lack of a universal definition of fairness or unfairness in the context of algorithmic bias. It will also specify the definition of an unfair outcome used in this thesis as it is required to be able to examine “unfair outcomes”.

Section 3 focuses on the first research question about biases causing unfair outcomes and how the biases end up in the algorithms. Section 4 focuses on the second research question and aims to showcase potential actions to combat unfair outcomes and whether they have already been utilized and the results and findings for future improvement.

Section 5 concludes the thesis and discusses the results. It also presents the implications (for research and practice), recommendations for future research and the limitations of the thesis.

The 3-D Dependable AI Framework (Ukanwa, 2024) is utilized throughout the research to tie together the findings from the multiple fields and present the possible solutions in a practical and easily understandable way. The framework is introduced in more detail in Section 2.

2 Key concepts for unfair outcomes caused by algorithmic bias

This section lays the groundwork and introduces the most important concepts for this thesis such as algorithmic bias and an unfair outcome. It also introduces the 3-D Dependable AI Framework.

2.1 Algorithmic bias

As we learned in the introduction, algorithmic bias can be divided into a multitude of different forms of bias. Generally, when talking about algorithmic bias it means prior information fed into the algorithm be it consciously or unconsciously that is used to make intelligent like decisions by the algorithm which in itself is not a problem but often these biases can contain "stereotyped" biases that cause negative outcomes (Caliskan et al., 2017). These biases are shown to have the potential to cause life-changing outcomes or for example make it harder for people to rise out of poverty (Wellner & Rothman, 2019; Zajko, 2021).

Xiang (2024) argues that AI algorithms are like a mirror that reflects the biased patterns in the training data it is created with. The algorithm is only as "good" as its training set and the people creating the algorithm, and the training data have a lot of influence over the future "fairness" of the algorithm. There are large differences in perspectives between different fields on what is important and what are the priorities when creating algorithms. There can also be differences in expertise among those involved in developing an algorithm. For example, the technical team may have expertise in areas like machine learning but may lack insight into the broader societal or ethical implications. Conversely, those with experience in ethical issues might not fully understand the technical aspects (Chen et al., 2023). This calls for interdisciplinary collaboration when creating algorithms and datasets used in them (Zajko, 2021).

2.2 Defining unfairness

Defining what is fair and what is unfair has proven to be a hard task at least in the context of algorithms and artificial intelligence (Bellamy et al., 2019; Grunwald, 2023) but it also seems like different versions of fairness can produce totally different outcomes (Bellamy et al., 2019) which indicates that finding the "correct" definition is really important and impacts the results of the research in a major way. Kalluri (2020) argues in her article that fairness should not even be the question that is asked because the definition of fairness can be molded to fit all kinds of situations where the reality might actually be discriminatory. In her view the better question to ask would be: "how is AI shifting power and who are the people wielding that power" (Kalluri, 2020, p. 169).

As this thesis focuses on unfair outcomes, a definition of that is going to be required. We will adopt this definition from a highly regarded journal from an article that has been cited by hundreds of papers. The definition goes as follows: "A system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate" (Friedman & Nissenbaum, 1996, p. 332). An example of this type of unfair outcome in the context of the thesis could be a citizen getting a wrong debt repayment claim from an AI-based and automated decision-making system, causing them financial distress and placing the onus of proof on the citizen (Rinta-Kahila et al., 2022).

2.3 3-D Dependable AI Framework

The 3-D Dependable AI Framework (Ukanwa, 2024) is a framework designed to gather together and bring light to different aspects of developing a fairer and more responsible artificial intelligence in all parts of its life cycle. The "3-D" comes from the 3 main stages of the life cycle of AI which are **design**, **develop** and **deploy**. These stages are then divided into smaller aspects of the stage which all need to be kept in mind during the life cycle of the artificial intelligence system. Figure 2 illustrates the framework.

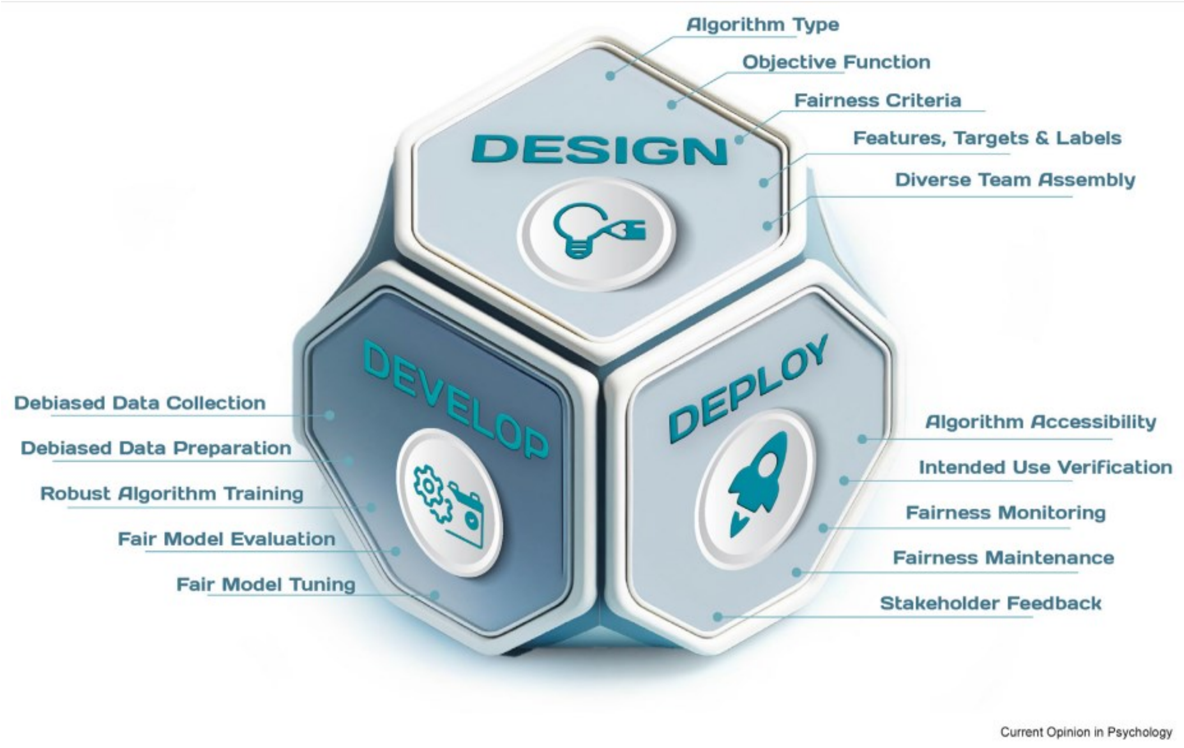

Figure 2. The 3-D Dependable AI Framework (Ukanwa, 2024)

The first part and the beginning of developing an AI is the **design** phase which includes crucial steps such as deciding the type of algorithm to be used and the functions of the AI and defining the desirable outcomes. This part is crucial especially from the ethical viewpoint as addressing harmful biases etched into the system becomes much harder in the later developmental stages of the AI life cycle (Zajko, 2021).

The second part is the **develop** phase which is the phase where the algorithm is trained and evaluated based on the choices made in the first phase. This is the part where training data is used to train the model which requires expertise and focus on the potential biases of that data (Ukanwa, 2024). It includes multiple rounds of running the data through the model and analyzing the results (Suresh & Guttag, 2021).

The third and last part of the framework is the **deploy** phase where the algorithm has been trained, tested and given the green light to be deployed for whatever its use case is (Ukanwa, 2024). It is crucial that the developers are monitoring the intended use and fairness of the system as there have been examples of things going very wrong very fast such as the case of a twitter chatbot “Tay” created by Microsoft which had to be shut down in under a day of its release when it started spewing hateful and racist comments (Bechmann & Bowker, 2019). This final part lasts as long as the system is running and requires continuous attention.

3 Unfair outcomes caused by algorithmic bias

A multitude of reasons exist for unfair outcomes caused by algorithmic bias and identifying them is key for preventing unfair outcomes in the future. Also, legal aspects might cause unfair outcomes. These will be presented next.

3.1 Differing sources of bias

There are different types of sources for biases that each require different types of approaches to be able to examine and understand them. Based on the extant literature, a broad distinction is made between two types of biases: cognitive bias and technical bias. Cognitive bias (also known as human bias) stems from the human brain whereas technical biases have more to do with the algorithm or the data used in it.

3.1.1 Cognitive biases

Cognitive biases include biases like for example confirmation bias, availability bias and anchoring bias (Rastogi et al., 2022). These biases may cause the human user to make skewed or incorrect decisions based on the potentially incomplete information or their prior beliefs and experiences. (Rastogi et al., 2022). This can happen both in the developmental phases of building and configuring the algorithm or in the phase where the algorithm is already in use, by for example a recruiter at work (Chin et al., 2023; Johnson, 2020).

Confirmation bias means seeking information or interpreting it in ways that is in line with the expectations, existing beliefs or the hypothesis in hand (Nickerson, 1998). Confirmation bias can manifest itself for example in social media usage where the algorithm might start feeding the user with things he already agrees with which can lead to “echo chambers” of separated groups where like-minded people isolate themselves from differing opinions, whether online or in real life. Coupled with the increase of misinformation available on the internet this can have harmful consequences (Zollo, 2019).

Availability bias is the tendency to evaluate probabilities or frequencies of events happening based on the availability of prior information regarding these things. Availability meaning the things that come up to the mind first and the fastest (Tversky & Kahneman, 1973). This can often mean that instead of utilizing the whole or at least larger proportion of the information available the most “available” things are only considered, and the possible decision is made with them leading to a potentially skewed decision.

Anchoring bias means as the name suggests anchoring to some type of input or information which then limits the willingness or capacity to explore alternative ideas or hypotheses. Effects of the anchoring bias are worsened if there is some type of time limit to the decision to be made. The “anchor” can come from a potentially flawed input or decision made by an AI that then steers the decision made by the human included in the decision-making process (Rastogi et al., 2022). In the accelerating work culture of today that sometimes requires very fast decisions made coupled with the helping hand of AI, anchoring bias can potentially have huge impacts in the decisions (Taeihagh, 2021). Availability bias and anchoring bias are similar in nature as both cause the person to underutilize the information available.

3.1.2 Technical biases

Technical biases represent the more technical sources for algorithmic bias. They are still often results of human bias but may require vastly different mechanisms to understand and prevent them. Technical biases may be divided into two different types of biases: data selection bias and label choice bias.

Data selection bias, also known as non-representative sampling, occurs when an unfair or unrepresentative data population is used in an algorithm, leading to outputs that do not accurately reflect the reality of the data population (Shah et al., 2019). “Debiased data collection” is also mentioned in the 3-D framework as an important characteristic of developing ethical AI (Ukanwa, 2024).

Label choice bias or annotation bias is defined by Obermeyer et al. (2019) as “the difference between some unobserved optimal prediction and the prediction of an algorithm trained on an observed label”. This type of bias can often happen unintentionally (Haliburton et al., 2024) as the reasons for these biases can happen because of cognitive biases such as confirmation bias (Chen et al., 2023). Nevertheless, label choice bias can skew the results of the algorithm and needs to be kept in mind during the developmental process of the algorithm. It is also emphasized in the “debiased data preparation” and “features, targets & labels” parts of the develop phase from the 3-D Dependable AI Framework (Ukanwa, 2024).

3.2 Autonomy of the system

Mittelstadt et al. (2016) states that algorithms and AI systems growing more complex also increases the autonomy of these systems. This means that the human using or supervising the system has less control over what is happening due to the complexity and amounts of code. Programs might also use external libraries that are sometimes called “black boxes” due to the lack of knowledge regarding their includings (Mittelstadt et al., 2016). This raises the question of who is in charge and who has the responsibility if something goes wrong with the system or the outcomes the system produces. It is especially hard to fully understand the more complex algorithms with learning capacities that also program themselves (Mikalef et al., 2022; Mittelstadt et al., 2016).

A clear example of the problems that arise from loss of human autonomy in the decision-making process is the case of the Australian government agency Centrelink (equivalent to Kansaneläkelaitos in Finland) automating its welfare overpayment detection and collection system (Rinta-Kahila et al., 2022). The government had had issues with overpayments of welfare and had decided to crack down on them. An algorithmic decision-making system coined “Robodebt” by the Australian media was put to place which due to multiple errors in the creation and implementation of the system resulted in “destructive effects” (Rinta-Kahila et al., 2022). The system resulted in thousands of incorrect debt recollections and ultimately ended with a 1,2 billion settlement lawsuit that paid back the incorrect debt recollections and other fees. This also resulted in damages to governmental reputation and finances (Rinta-Kahila et al., 2022). The system was put to use even when employees and experts had raised concerns during the developmental process which were ignored largely due to top management’s tunnel vision and negative attitudes towards welfare services and their receivers (Rinta-Kahila et al., 2022).

3.3 Legal and regulatory shortcomings

So far we have talked about algorithmic bias from the perspective of the companies and the users but there is also another big player which is the governments. The governments have a big role in determining the playing field and the rules in which the companies and people operate in. With the complex and global nature of the algorithmic systems that may impact people all around the world, the task of regulating and policing the use of AI is not an easy one (Taeihagh, 2021). Issues like loss of privacy have been a topic for years now and there seems to be a lot of people unhappy with the current level of privacy protection with countries like Japan where 80% of respondents consider the current level of privacy protection to be insufficient. Another similar survey finds that 3 in 4 of Chinese people are concerned about privacy. (Mazurek & Małagocka, 2019)

3.3.1 Self-governance and lack of standardized practices

Currently there seems to be a lack of standardized industry practices which would set clear frameworks for testing and implementing the systems and reporting the progress of doing so (Mitchell et al., 2019; Taeihagh, 2021). This can lead to every company having their own procedures with varying degrees of importance associated to for example ethical considerations, this coupled with the somewhat self-governing nature of the industry right now can have adverse impacts (Hagendorff, 2020; Mitchell et al., 2019). Research shows that current guiding ethical guidelines have almost no impact on the behavior of professionals in the tech community and as the companies developing and maintaining the systems are driven by an economic logic there seems to be no encouragement for workers to raise ethical concerns (Hagendorff, 2020).

There seems to be also a problem of competing for global political power between China, the United States of America and the EU which diminishes governmental incentive to implement strong regulations surrounding artificial intelligence (Hagendorff, 2020). Artificial intelligence has major implications for example in the military industry that is a big part of the global competition for power (Hagendorff, 2020).

3.3.2 Speed of progress

AI and computer systems in general have been developing at an unprecedented rate which has led to regulatory challenges to keep up with the development (Taeihagh, 2021; Wirtz et al., 2020). Coupled with the huge informational and resource advantages of large companies' regulations so far have largely been made with the "soft law" approach which includes a lot of self-governance and "non-binding norms and techniques" which has given the companies a lot of independence to determine the systems and their requirements (Hagendorff, 2020).

Independency is good for the development of the algorithms and AI systems that use them, but it can mean more problems for the users and people in ways that have been discussed in this thesis so far. There needs to be a balance between protecting the people that these systems affect in negative ways and possibility for development of systems that can have positive effects on society (Taeihagh, 2021; Wirtz et al., 2020). This is of course a political question which has no objective right answer but is also largely shaped by the current political and moral atmosphere of the world (Wirtz et al., 2020).

4 Preventing unfair outcomes caused by algorithmic bias

This section showcases possible ways of preventing unfair outcomes caused by algorithmic bias within the framework of the 3-D Dependable AI Framework. It is divided into 3 subsections that each portray a part of the 3 stages of the AI development life cycle (Ukanwa, 2024).

4.1 Design phase

The design phase of the framework includes 5 areas that are the following: algorithm type, objective function, fairness criteria, features, targets & labels and diverse team assembly (Ukanwa, 2024). This is the part where human developers still retain most of the control, so it is a crucial step that creates the playing field for the rest of the life cycle of the AI and partly determines the future possibilities for interventions in the later stages.

Algorithm type selection and the objective function of the algorithm start the process and have the potential already to reflect priorities, opinions and preconceptions that may introduce algorithmic bias (Ukanwa, 2024). It is important for the developers of the algorithm to keep these things in mind. An example of the importance of algorithm type selection and defining the objective function can be found from the health care sector where a widely used and esteemed algorithm was shown to discriminate against black people (Obermeyer et al., 2019). This highlights the need to critically evaluate even existing algorithm types with a widely accepted objective function.

Choosing the features, labels and targets has a major impact on the outcomes produced and on what basis they are made on (Johnson, 2020; Obermeyer et al., 2019; Ukanwa, 2024; Wellner & Rothman, 2019). Features are the independent variables that the algorithm uses to make predictions or classifications; they can be for example sensitive attributes such as race, gender or income (Obermeyer et al., 2019) that have been shown to produce discriminatory outcomes (Wellner & Rothman, 2019). Labels and targets are the dependent variables that the algorithm is predicting or classifying, for example categorizing job or loan applicants (Johnson, 2020). Avoiding the use of sensitive attributes, such as gender or race, has been shown to be insufficient on its own as the algorithms can learn to utilize these attributes indirectly such as utilizing zip codes in much of the same way as race because zip codes have been shown to correlate with peoples' races (Johnson, 2020; Wellner & Rothman, 2019). Using features is detrimental to the accuracy of the algorithm and currently there seems to be an intricate trade-off between dependence on features and accuracy judgements (Johnson, 2020). This means that the accuracy of the model weakens the less these features are used. As many of the features have been shown to produce discriminatory results it seems like the problem is not so much in the features but in the discriminatory nature of society itself. Zajko (2021) suggests that every problem cannot be solved by designing the system better but by actually fixing the underlying societal inequalities such as gender underrepresentation in certain fields of work that have been shown to exist partly because of explicit and implicit biases (Johnson, 2020; Suresh & Guttag, 2021; Wellner & Rothman, 2019; Zajko, 2021). Tweaking the dependent variables has been shown to reduce biases, in the healthcare sector an algorithm that was considered to be an industry leader had a dependent variable that was predicting future cost which was producing discriminatory results to black people. Instead of just focusing on the cost, health prediction was added to the dependent variable which resulted in a major reduction of 84% in bias (Obermeyer et al., 2019).

Developing a fair AI, free of algorithmic bias requires defining when the algorithm is fair versus biased (Ukanwa, 2024). Unfortunately, there is no universal definition of fairness, and it is dependent on, for example cultural values and social norms that can differ between people from different backgrounds (Chen et al., 2023; Ukanwa, 2024). As such there is a need to get a broad view from people that are going to be affected by the system to be able to take their needs and particularities into account (Chin et al., 2023; Wellner & Rothman, 2019).

In general, a diverse set of viewpoints makes it possible to see the issues more broadly and to gain new insights. Aside from being an aspect of preventing algorithmic bias, diverse development team assembly is shown to come with a multitude of performance benefits, for example in problem-solving, innovation and creativity (Van Knippenberg et al., 2020). Diversity should come at least in the forms of ethnicity and gender (Wellner & Rothman, 2019) improving the capability of the developers to sense the potential issues before they manifest into algorithmic bias (Chin et al., 2023; Turner Lee, 2018).

4.2 Develop phase

The second phase in the framework is the develop phase in which the algorithm is trained and evaluated based on the data gathered and prepared for the algorithm. It includes 5 areas that are: debiased data collection, debiased data preparation, robust algorithm training, fair model evaluation and fair model tuning (Ukanwa, 2024). As discussed already in Section 3.1, data selection can manifest multiple types of biases from the cognitive biases to the technical biases. Shah et al. (2019) introduce multiple ways for dealing with different types of biases like label (choice) bias, selection bias, overamplification and semantic bias. These countermeasures will now be briefly introduced.

4.2.1 Data selection and preparation

Label bias, as discussed in Section 3.1, can come from the biases of the people annotating the data or from differences between different annotators (Haliburton et al., 2024; Obermeyer et al., 2019). Bayesian annotation models can be used to embrace and examine the differences by assigning confidence scores to each annotator and reweighting them based on the results (Shah et al., 2019). This has led to major reductions in label bias.

Selection bias is the misrepresentation of the actual population group. The easiest way to prevent selection bias is re-stratification where the overly represented subgroups' samples are reduced to reflect the ideal distribution more accurately. However, reducing the total sample size can have its own problems. Other ways include stratified sampling techniques, reweighting or poststratifying (Shah et al., 2019). Stratification sampling in the context of selection bias can mean for example grouping the data into subgroups based on relevant characteristics, such as age or gender, to be able to compare and match the data to the target populations (Filho et al., 2015; Shah et al., 2019).

Overamplification means the algorithm picking up on small differences between attributes and amplifying their predictive values way out of proportion compared to the actual differences (Shah et al., 2019). Multiple ways exist to bring down the weight of the instances in the algorithm for example using synthetic matched distributions which alter the distributions to decrease overamplification (Shah et al., 2019).

Semantic bias can manifest for example in the form of a word such as doctor being way more associated with males compared to females in the algorithm (Johnson, 2020; Shah et al., 2019). Techniques to prevent semantic bias can be generally divided into hard de-biasing and soft de-biasing where hard de-biasing completely removes semantic bias but has the potential to introduce other problems such as reducing the classifiers' performance. Soft de-biasing on the other hand is not as effective but avoids side effects (Bolukbasi et al., 2016; Shah et al., 2019).

4.2.2 Model training and evaluation

After the data is collected and prepared it is used to train the model which includes multiple phases of training the model, evaluating the outcomes and adjusting the model based on the results. Evaluating the model's performance and outcomes is crucial when developing ethical artificial intelligence

(Ukanwa, 2024). This is also the part where it is still possible to notice and effectively prevent bias arising from the data collection and preparation.

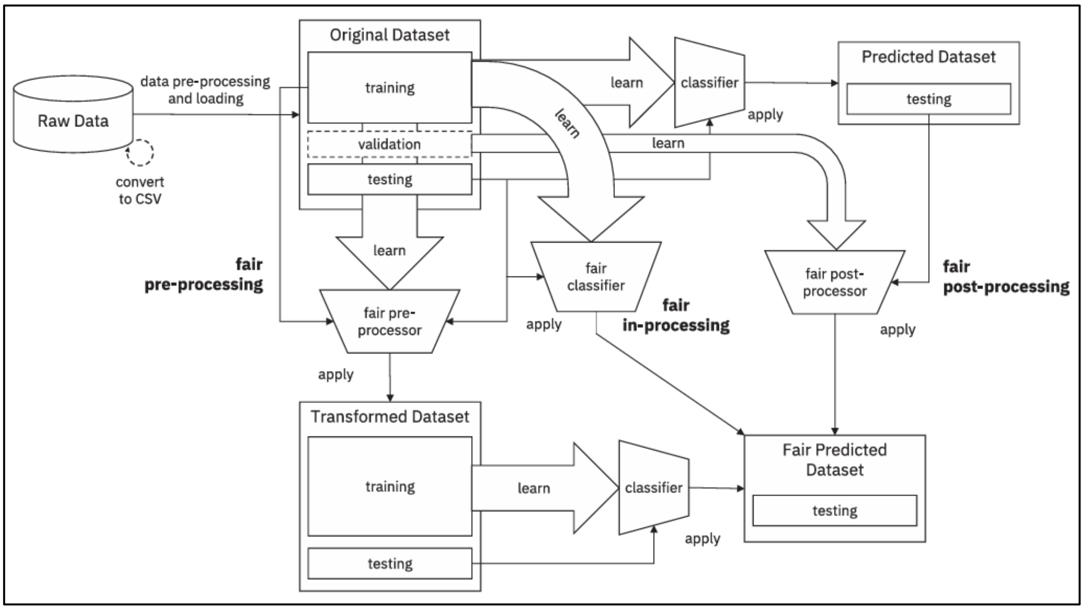

Figure 3. Fairness pipeline (Bellamy et al., 2019)

Bellamy et al. (2019) introduce in their paper an AI Fairness 360 Python toolkit designed to help developers evaluate the fairness of their algorithms. It includes over 70 detection metrics for biases, 9 bias mitigation algorithms and guidance in the form of tutorials and notebooks for developers. Figure 3 showcases the "fairness pipeline" which is a generic example of using the toolkit. Data and the algorithm can be evaluated in three phases: pre-processing, in-processing and post-processing. All three phases are available for developers who have access to everything. (Bellamy et al., 2019)

Evaluation of the results and whether the algorithm developed is ethical is still largely dependent on the values and beliefs of the people developing the algorithm. This calls for large and active involvement of the target groups being affected by the algorithm, especially those who have been historically marginalized (Chin et al., 2023; Kalluri, 2020; Wellner & Rothman, 2019). If the goal is an AI system with positive impact, devoid of unfair outcomes, the people that are most likely going to be affected negatively need to be heard and their needs must be considered as AI systems by default are more prone to reproduce underlying societal inequalities rather than transforming them (Wellner & Rothman, 2019; Zajko, 2021).

4.3 Deploy phase

The deploy phase is the last one on the framework and it includes 5 areas that are: algorithm accessibility, intended use verification, fairness monitoring, fairness maintenance and stakeholder feedback (Ukanwa, 2024). This is the part where the system is deployed for whatever its intended use is and things like roll-out strategies are created alongside forming monitoring and maintenance processes (Ukanwa, 2024). Even if the system was created “as good as possible” with the framework or something similar in mind there is still a possibility of the system negatively interacting with something that could not have been noticed during the design and develop phase (Zajko, 2021). This is the perpetual nature of developing and maintaining artificial intelligence which means on-going monitoring and involvement of users to prevent for example unfair outcomes (Ukanwa, 2024; Wellner & Rothman, 2019).

4.3.1 Algorithm accessibility

Algorithm accessibility in the current context means how much of the processes and data are available for cohorts outside of the developers and owners of the algorithms. This can mean for example researchers who might want to study the possible algorithmic biases but currently they are often met with the problem of most of the data and information needed, being hidden or not available thus hindering the possibility of conducting empirical investigations (Obermeyer et al., 2019). There often are multiple conflicting interests regarding who gets access to what kind of data about the algorithms. Companies are inclined to keep as much information out of potential competitors’ hands, data subjects are interested in what kind of information about them is being processed and how, and researchers are interested in for example the underlying processes and ethical implications of the algorithms (Mittelstadt et al., 2016). There is also clearly an unequal power dynamic between companies and data subjects where the companies producing and maintaining the algorithms have informational availability and ability at levels completely unattainable to singular data subjects (Mittelstadt et al., 2016).

Algorithmic accessibility is of the utmost importance for preventing unfair outcomes because it increases accountability and information on how the algorithms are working and on what kind of basis the decisions are made (Keats et al., 2014; Mittelstadt et al., 2016; Obermeyer et al., 2019).

4.3.2 Intended use and fairness monitoring

Artificial intelligence has the potential to shape individual aspects of our lives and even society, but these things also shape how AI systems function, thus intended use and fairness monitoring are a continuous process that should be done for as long as the system is active (Mikalef et al., 2022; Ukanwa, 2024; Wellner & Rothman, 2019). Monitoring also requires the ability to admit that the system is not working as intended or that it is causing harm. The cautionary example of "Robodebt" (discussed in Section 3.2) highlights the importance of critically examining processes surrounding intended use and fairness monitoring especially when the automated decision-making system is also delivering intended effects. The intended effects such as increased revenue in the case of Robodebt caused the decision-makers to turn a blind eye to the negative effects and escalate commitment to the project (Rinta-Kahila et al., 2022). This surprisingly common phenomenon of "escalation of commitment", where a "bad" project like Robodebt is allowed to continue (Keil et al., 2000) for reasons like amount of money spent or personal career development, needs to be kept in mind when considering monitoring processes (Rinta-Kahila et al., 2022).

Stakeholder feedback plays an important role in the prevention of unfair outcomes but also requires efforts from the entity implementing the system (Chen et al., 2023). Accountability requires transparency and transparency requires comprehensibility (Mikalef et al., 2022; Mittelstadt et al., 2016). Transparency and comprehensibility introduce potential problems, and effort needs to the entity implementing the system and the system itself, like potential privacy issues and the highly complex nature of algorithmic decision making which can be hard or impossible to understand sometimes even to the developers themselves (Mittelstadt et al., 2016). Nevertheless, these topics need to be considered for the system to be responsible (Mikalef et al., 2022). Arguments have also been made that the more the system is affecting certain people's lives the more those specific people should have a say in the development and maintenance of that system (Kalluri, 2020; Wellner & Rothman, 2019; Zajko, 2021).

Table 1 summarizes the findings of the thesis as per the three areas of the 3-D Dependable AI Framework (Ukanwa, 2024).

Table 1. Findings for areas of the framework

Life cycle stage		Findings
Develop	Algorithm type	Existing algorithm types for the field in question need to be evaluated and if they do not fit the standards for fairness, a new fair algorithm needs to be created
	Objective function	Sometimes better (cost-)efficiency means more unfair outcomes. Determining the function requires fairness as an important factor.
	Fairness criteria	As there is no universal definition of fairness, a diverse set of viewpoints that includes future stakeholders is required to establish effective fairness criteria.
	Features, targets & labels	Features determine the outcomes that are defined by targets and labels. Target and label choice has a large impact on bias so deciding it requires lots of attention.
	Diverse team assembly	A diverse range of viewpoints allows for a broader understanding of issues and helps discover new insights.
Design	Debiased data collection	Data collection may require for example grouping the data into subgroups based on relevant charasteristics to be able to match the data to the target populations.
	Debiased data preparation	Involves issues like overamplification, where algorithms exaggerate small attribute differences, and semantic bias, such as gendered associations with words.
	Robust algorithm training	The algorithm goes trough multiple phases of training where its outcomes are evaluated and adjusted. Multiple toolkits exist to help developers train the algorithms.
	Fair model evaluation	Toolkits like the "AI Fairness 360 Python toolkit" exist to help developers evaluate the fairness of the algorithm trough detection metrics and bias mitigation algorithms.
	Fair model tuning	Tuning decisions are largely dependent on the values of the developers calling for large involvement of target groups affected by the algorithm.
Deploy	Algorithm accessibility	Accessibility is of the utmost importance because it increases accountability and information on how the algorithms are working and on what kind of basis.
	Intended use verification	AI is also shaped by the society around it and thus can react to it in ways that are not intended. This requires active monitoring of the outcomes.
	Fairness monitoring	Even if the system is working as intended, it can produce unfair outcomes. Stopping or fixing the system requires negating phenomenons like "escalation of commitment".
	Fairness maintenance	Preventing escalation of commitment requires measures to ensure that factors such as the amount of money already invested or career development do not contribute to it.
	Stakeholder feedback	Feedback necessitates transparency and comprehensibility which requires efforts from the developers due to the complex nature of algorithms.

5 Conclusions

This thesis examined unfair outcomes caused by algorithmic bias and the socioethical aspects surrounding it. It presented some findings on how unfair outcomes caused by algorithmic bias could be prevented and in general how to build more ethical AI with the help of the 3-D Dependable AI Framework. The issues surrounding the topic are extremely complex and multidisciplinary and are developing every day with the rapid development of AI systems.

5.1 Limitations

The length of a bachelor's thesis limited how extensively the possible methods for preventing unfair outcomes could be explored. Some of the solutions were also extremely complex, requiring serious expertise and years of experience in the fields of algorithmic studies and psychology for example.

Focusing on the ethical topics regarding AI systems and aiming towards responsible systems is a fairly new topic in research. A lot of the past studies have focused on the positive and beneficial aspects (Mikalef et al., 2022) and especially from the views of the companies and governments creating the systems (Kalluri, 2020). Thus, there are still lots of unknowns and issues that have not been discovered, especially from the viewpoints of the data subjects, users of the systems and the sections of population who have historically been marginalized (Chin et al., 2023; Kalluri, 2020).

5.2 Implications for practice

Based on the thesis, the 3-D Dependable AI Framework appears to be a capable tool for the development and examination of AI, and it, along with similar frameworks, should be used and considered when working with AI. Alongside the frameworks, toolkits like the AI Fairness 360 toolkit should also be utilized.

Developers and organizations utilizing AI in their systems should be aware of the ethical implications and possible dangers of AI usage. Organizations should strive towards organizational structures that prevent situations where escalation of commitment causes a harmful system to be maintained due to reasons of personal career development or fear of consequences like getting fired. The cautionary example of Robodebt is a clear example of the dangers of escalation of commitment and highlights the need for organizations to have systems and structures in place that prevent similar things from happening.

On a governmental level, the lack of regulations seems to be an issue that needs attention especially outside of the EU. Regulations and guidelines have been deprioritized, partly due to the accelerating competition in AI development between geopolitical superpowers such as China and the USA. Issues that individual users might face should not be overlooked or deprioritized in the pursuit of a competitive advantage.

5.3 Future research

It is clear from the findings and issues brought forward by this thesis that the topics need more interdisciplinary research and collaboration between the multiple fields that revolve around the topic. Currently there seems to be information asymmetry between different fields that hinders the development of ethical AI. Future research should utilize and combine experts and research from the variety of fields surrounding the issues of development of AI, with ethics prioritized as a central focus.

Future research should also look into the issue of finding a balance between allowing for the development of AI and upholding rights of individuals, for example in relation to privacy. It should assess how the positive effects of AI could be harnessed while simultaneously minimizing the negative and destructive effects. This should be evaluated on a governmental level regarding regulations and what are the best practices but also at an organizational level.

References

Bechmann, A., & Bowker, G. C. (2019). Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media. *Big Data and Society*, *6*(1). https://doi.org/10.1177/2053951718819569

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. *IBM Journal of Research and Development*, *63*(4–5). https://doi.org/10.1147/JRD.2019.2942287

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*. http://arxiv.org/abs/1607.06520

Buolamwini, J. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research*, *81*, 1–15.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. In *Science* (Vol. 356). https://www.science.org

Chen, Y., Clayton, E. W., Novak, L. L., Anders, S., & Malin, B. (2023). Human-Centered Design to Address Biases in Artificial Intelligence. In *Journal of Medical Internet Research* (Vol. 25). JMIR Publications Inc. https://doi.org/10.2196/43251

Chin, M. H., Afsar-Manesh, N., Bierman, A. S., Chang, C., Colón-Rodríguez, C. J., Dullabh, P., Duran, D. G., Fair, M., Hernandez-Boussard, T., Hightower, M., Jain, A., Jordan, W. B., Konya, S., Moore, R. H., Moore, T. T., Rodriguez, R., Shaheen, G., Snyder, L. P., Srinivasan, M., … Ohno-Machado, L. (2023). Guiding Principles to Address the Impact of Algorithm Bias on Racial and Ethnic Disparities in Health and Health Care. In *JAMA Network Open* (Vol. 6, Issue 12, p. E2345050). American Medical Association. https://doi.org/10.1001/jamanetworkopen.2023.45050

Filho, R. M., Almeida, J. M., & Pappa, G. L. (2015). *Twitter Population Sample Bias and its impact on predictive outcomes: a case study on elections*. http://www.academia.org.br/

Friedman, B., & Nissenbaum, H. (1996). Bias in Computer Systems. *ACM Transactions on Information Systems*.

Grunwald, B. (2023). Racial Bias in Criminal Records. *Journal of Quantitative Criminology*, *40*, 489–531. https://doi.org/10.1007/s10940-023-09575-y

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Haliburton, L., Ghebremedhin, S., Welsch, R., Schmidt, A., & Mayer, S. (2024). Investigating Labeler Bias in Face Annotation for Machine Learning. *Frontiers in Artificial Intelligence and Applications* , *386*, 145–161. https://doi.org/10.3233/FAIA240191

Johnson, G. M. (2020). Algorithmic bias: on the implicit biases of social technology. *Synthese*, *198*(10), 9941–9961. https://doi.org/10.1007/s11229-020-02696-y

Kalluri, P. (2020). Don't ask if AI is good or fair, ask how it shifts power. *Nature*. https://doi.org/10.1038/d41586-020-02003-2

Keats, D., Pasquale, C. F., Citron, D. K., & Pasquale, F. (2014). *The Scored Society: Due Process for Automated Predictions*. http://ssrn.com/abstract=2376209

Keil, M., Tan, B. C. Y., Wei, K.-K., Saarinen, T., Tuunainen, V., & Wassenaar, A. (2000). A Cross-Cultural Study on Escalation of Commitment Behavior in Software Projects. In *Source: MIS Quarterly* (Vol. 24, Issue 2).

Mazurek, G., & Małagocka, K. (2019). Perception of privacy and data protection in the context of the development of artificial intelligence. *Journal of Management Analytics*, *6*(4), 344–364. https://doi.org/10.1080/23270012.2019.1671243

Mikalef, P., Conboy, K., Lundström, J. E., & Popovič, A. (2022). Thinking responsibly about responsible AI and 'the dark side' of AI. In *European Journal of Information Systems* (Vol. 31, Issue 3, pp. 257–268). Taylor and Francis Ltd. https://doi.org/10.1080/0960085X.2022.2026621

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 220–229. https://doi.org/10.1145/3287560.3287596

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data and Society*, *3*(2). https://doi.org/10.1177/2053951716679679

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. In *Review of General Psychology* (Vol. 2, Issue 2).

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/SCIENCE.AAX2342

Rastogi, C., Zhang, Y., Wei, D., Varshney, K. R., Dhurandhar, A., & Tomsett, R. (2022). Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW1). https://doi.org/10.1145/3512930

Rinta-Kahila, T., Someh, I., Gillespie, N., Indulska, M., & Gregor, S. (2022). Algorithmic decision-making and system destructiveness: A case of automatic debt recovery. *European Journal of Information Systems*, *31*(3), 313–338. https://doi.org/10.1080/0960085X.2021.1960905

Shah, D., Schwartz, H. A., & Hovy, D. (2019). *Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview*. https://doi.org/10.18653/v1/2020.acl-main.468

Suresh, H., & Guttag, J. (2021, October 5). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3465416.3483305

Taeihagh, A. (2021). Governance of artificial intelligence. *Policy and Society*, *40*(2), 137–157. https://doi.org/10.1080/14494035.2021.1928377

Turner Lee, N. (2018). Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society*, *16*(3), 252–260. https://doi.org/10.1108/JICES-06-2018-0056

Tversky, A., & Kahneman, D. (1973). Availability: A Heuristic for Judging Frequency and Probability. In *COGNlTTIVE PSYCHOLOGY* (Vol. 5).

Ukanwa, K. (2024). Algorithmic bias: Social science research integration through the 3-D Dependable AI Framework. In *Current Opinion in Psychology* (Vol. 58). Elsevier B.V. https://doi.org/10.1016/j.copsyc.2024.101836

Van Knippenberg, D., Nishii, L. H., & Dwertmann, D. J. G. (2020). *Synergy from diversity: Managing team diversity to enhance performance*.

Wellner, G., & Rothman, T. (2019). *Feminist AI: Can We Expect Our AI Systems to Become Feminist?* https://doi.org/10.1007/s13347-019-00352-z

Wirtz, B. W., Weyerer, J. C., & Sturm, B. J. (2020). The Dark Sides of Artificial Intelligence: An Integrated AI Governance Framework for Public Administration. *International Journal of Public Administration*, *43*(9), 818–829. https://doi.org/10.1080/01900692.2020.1749851

Xiang, A. (2024). Mirror, Mirror, on the Wall, Who's the Fairest of Them All? *Daedalus*, *153*(1), 250–267. https://doi.org/10.1162/daed_a_02058

Zajko, M. (2021). *Conservative AI and social inequality: conceptualizing alternatives to bias through social theory*. *36*, 1047–1056. https://doi.org/10.1007/s00146-021-01153-9

Zollo, F. (2019). Dealing with digital misinformation: a polarised context of narratives and tribes. *EFSA Journal*, *17*(S1). https://doi.org/10.2903/j.efsa.2019.e170720