Exploring Bias Against Women in Artificial Intelligence

Practitioners' Views on Systems of Discrimination

Cathrine Kieu Trang Bui

Master’s Thesis
Programming and Networks
60 credits

Department of Informatics
The Faculty of Mathematics and Natural Sciences

UNIVERSITY OF OSLO

June / 2021

Exploring Bias Against Women in Artificial Intelligence

2021

Exploring Bias Against Women in Artificial Intelligence

http://www.duo.uio.no/

Trykk: Reprosentralen, Universitetet i Oslo

Abstract

Background: AI systems increase in popularity and widely implemented in many areas. Media and literature have reported numerous incidents of discriminating AI systems. Literature has identified several causes and solutions to gender bias in AI, and many institutions have published ethics guidelines. However, previous research has not studied the perspectives and practices of practitioners in AI.

Aim: This thesis explores what perspectives practitioners in AI in Norway have on gender bias in AI by investigating their understanding of technology; how gender bias enters AI systems; and what practices they have in place to detect and address gender bias in AI.

Method: Qualitative multiple case studies were conducted. This study interviewed 13 practitioners in the AI field in Norway. Thematic analysis was used to analyze the interviews.

Findings: Practitioners have implemented few practices, most do not use any ethics guidelines, and they delegate responsibilities to other entities. The informants could only identify a few of the entry points of gender bias mentioned by literature, such as biased data, human bias, and a lack of diverse perspectives. The informants with at least one marginalized identity had more knowledge and practices to address gender bias in AI. They were able to identify more systemic causes and higher-impact levers of intervention.

Conclusion: AI practitioners have inherited assumptions and beliefs from predecessors in the AI field on how distancing oneself from one's work achieves neutral objectivity. These beliefs have a significant influence on practitioners' understanding of technology, and as a result, few ethics practices are in place. These assumptions conflate their grasp of what causes gender bias in AI into a technical problem because they underestimate the effects of power. The practitioners see biased data as the main cause, but data is never neutral because no dataset is equally fair for everyone. The practitioners' belief that there exists a form of fairness that will always be correct for everyone at all times without considering the context enables biases to enter AI systems. The AI field needs to examine what technical heritage and taken-for-granted beliefs negatively impact research and practices on gender bias in AI. This study recommends a paradigm shift in practitioners from imagined objectivity to a critical, intersectional perspective that empowers, includes, and creates justice for disadvantaged groups. Inclusion of marginalized perspectives is crucial, and hiring practices should change to increase diversity by training disadvantaged groups in AI.

Acknowledgements

This thesis has been the most difficult project I have ever completed and I am indebted to many who have supported me along the way. A big thank you to:

My supervisor Maja Van Der Velden for teaching me how to critically assess the world, and taking women's issues seriously, thus enabling me to write a thesis about women's issues.

My initial supervisor, Kyrre Begnum at OsloMet, for teaching me about myself and being my academic personal trained in times of distress.

My other supervisor Yngve Lindsjørn, for having my back through the years.

Additional expressions of gratitude are expressed to supervisor Maja and Andrea Gasparini's organization of the Sustainability and Design Lab, and to its wonderful members and other peers at IFI's seventh floor. Also, thank you to my supportive friends in the trenches, including, but not limited to: Matthew Smart, Hadiya Firdaus, Rannveig Skjerve, Miya Perry, and others.

Thank you to my lawyer Thomas Benestad for being my hero this last year.

And lastly, perhaps one of the biggest expressions of gratitude to Lars Lyngstad Sund for supporting me, feeding me, and being patient with me this last year.

Contents

List of Tables

List of Figures

1 Introduction

De Spiegeleire, Maas & Sweijs report that Artificial intelligence (AI) has made a comeback after the previous AI winter of 1987-1993 thanks to advances in computing power. Moore's Law catapulted the world into an "AI Revolution". AI's lofty promises include improving people's lives, society and the economy. AI is projected to give us safer traffic, energy efficiency, more precise surgeries, and more efficient public administration. (De Spiegeleire, Maas, & Sweijs, 2017; European Commission, 2019)

However, AI is not simply a promise, it is a fact. Some researchers argue that the public mistakenly believe that the age of AI is about to descend on us when in reality it has been here for years (Hunter, Sheppard, Karlén, & Balieiro, 2018). When difficult tasks become solvable, they are considered simple, and Artificial Intelligence is no longer considered as intelligent because the task is "simple". Hunter et. al refer to this shift as the "AI effect", which contributes to the illusion of the world being on the cusp of the AI era when in reality we are already using it (2018). AI is here, and its problems are currently affecting our society. AI affects our lives and the decisions we make, from what ads we see (Lambrecht & Tucker, 2019) to who gets out of prison (Angwin, Larson, Mattu, & Kirchner, 2016).

AI has great power to process data and generate decisions that have important impacts. Massive data power can process billions of pictures and text and surpass the efficiency of humans in simple tasks, such as facial recognition or translations. This pronounced superpower is predicted to solve problems humans cannot solve, such as the cure to cancer. This great power of AI comes with great responsibility.

Because of the power and reach of AI, the bias in it might become a great, powerful, and far-reaching problem. Inspired by 80,000 Hours, I wanted to address the issue of ethics in AI. 80,000 Hours is a non-profit organization that have dedicated their work to help corporate professionals change career paths to solve the most important problems in the world (80,000 Hours, n.d.-a). They have made a list of the biggest and most pressing problems of our time that needs to be prioritized and "Positively shaping the development of artificial intelligence" is the problem that has been chosen for this thesis (80,000 Hours, n.d.-b).

AI has a problem with bias because it is trained on data that can be biased and there have been several incidences where AI has not worked as intended. This bias can be gendered and affect women more than men. For instance, Microsoft's Twitter bot that turned Nazi, misogynistic, and racist from user inputs (Vincent, 2016) or the resume filtering AI at Amazon that filtered out all the women (Dastin, 2018). Studies have shown that AI classification of dark-skinned women does not work as well as for white men (Buolamwini & Gebru, 2018) and that voice assistants do not work as well for female voices (Tatman, 2017).

Some of these problematic examples sparked my interest in gender bias. Reading about these made me worried about AI's the future repercussions on society. My interest in gender bias in AI started in 2017 when I was preparing a talk for an event called Girls and Technology. I wanted to tell high school girls about the importance of their contribution to the STEM field because the consequences of a male-dominated technology workforce can be dire. I found countless examples of technologies that were designed for men and were ill-fitted for women. This interest was then amplified as I grew angrier after each chapter I read of the books Invisible Women (Perez, 2019) and Weapons of Math Destruction (O'Neil, 2016).

The reach of AI systems are limitless as software is easily deployed to the entire world's population that is connected to the world wide web. Open-source projects can be copied and forked with an instant click of a button. If one were to create a project based on someone else's code this would mean the project would also inherit its flaws, biases, and bugs. The common practice of forking projects and using open-source AI algorithms would mean that any bias that might exist in the original code would seep into countless other projects and products. Such as image recognition algorithms that are trained on data that have more samples of white male faces, and then struggle to recognize faces that are not white and male (Buolamwini, 2018; Buolamwini & Gebru, 2018).

How AI works inside its black box is a mystery that some are trying to unfold (Abdul, Vermeulen, Wang, Lim, & Kankanhalli, 2018), while many algorithms are kept secret as they are considered proprietary software. The detrimental effects of the cases mentioned above were not discovered until they had already been put to use. One of the challenges related to gender bias in AI is the lack of research and understanding of gender bias in AI product and service development.

Most AI developers and researchers are male. Some estimate the number of women in AI to be 26% ("Global Gender Gap Report 2020," 2019). Some people theorize that AI has a gender bias because of the skewed gender ratio (S. M. West, Whittaker, & Crawford, 2019). The status of this gender disparity will be elaborated in a later chapter.

What happens if the issue of gender discrimination in AI continues? What is the future of AI and feminism? Will oppressive AI systems undo decades of feminist progress as Leavy suggests

(2018)? Is the problem that AI developers and researchers are not aware of this issue, that they don't care, or that they don’t have ethics on the agenda? This thesis aims to provide a more organized overview of the issue of gender bias in AI, and suggestions for what to do about it. This thesis hopes to create knowledge that can aid AI communities in monitoring and checking for gender bias.

The main activity to address this problem is conducting interviews with AI experts to investigate the issue. The result of this thesis will be a model of how gender bias is introduced and a comparison of different causes and solutions.

1.1 Research Questions

Several research projects referenced in chapter 2 have investigated the problem with gender bias in AI. However, it is difficult to find any research on what the people who are creating the AIs are doing. How do they work? What are their workflows? Are they aware of the issues of gender bias? What do they do about the issues? Does anyone use the ethical guidelines that exist?

Based on what was found in the subsequent literature review, this thesis will be investigating the following research questions:

Main research question:
What are the main perspectives on gender bias in AI among AI practitioners in Norway?

Sub-question 1:
What understandings of technology are found among AI practitioners?

Sub-question 2:
How does gender bias enter an AI system?

Sub-question 3:
What practices are in place to detect and address gender bias in AI?

The practitioners' perspectives on gender bias in AI are relevant to explore because they can indicate the status of awareness and progress on addressing gender bias in AI. Their understandings of technology might affect their perceptions on how AI enters an AI system, which are related to their perspectives on gender bias in AI. However, regardless of how they view the issue, what they actually *do* about gender bias in AI might reveal more information about what they really think. The three sub-questions are therefore used to answer the main research question.

1.2 Definitions

AI Practitioner

This thesis uses the term *AI practitioner* as an umbrella term that refers to AI developers, AI researchers, designers, and others who work with creating or developing AI systems and services.

Artificial Intelligence (AI)

This thesis uses EU’s definition of artificial intelligence: "Artificial intelligence (AI) refers to systems that show intelligent behavior: by analyzing their environment they can perform various tasks with some degree of autonomy to achieve specific goals." (European Commission, 2019)

Therefore this thesis refers to artificial intelligence as algorithms and systems that perform tasks or make decisions based on input data. This includes algorithms such as machine learning and deep learning. Examples of such algorithms are voice assistants such as iPhone’s Siri, facial recognition systems such as automated passport controls in airports, or algorithmic decision systems such as automated processing of job applications or student loan applications.

Bias

The Merriam-Webster Dictionary defines bias as: "[A]n inclination of temperament or outlook; *especially*: a personal and sometimes unreasoned judgment: prejudice" ("Bias," 2020). Bias in this thesis refers to the questions if and how humans and AIs discriminate against a group or groups of people. The terms bias, gender bias, racial bias, and class bias in this thesis refers to the practice of unfairly discriminating against an individual based on certain traits.

These terms are not to be confused with the term bias as it is defined in the field of statistics. The term *statistical bias* is used when referring to that definition. Statistical bias refers to when an algorithm does not accurately represent the data; in this sense, an algorithm *should* include any gender bias that is present in the data to not be statistically biased. "Gender bias" in this thesis is sometimes used synonymously with "gender discrimination". "Gender bias in AI" will sometimes be abbreviated as GBAI.

Big Tech

The term Big Tech refers to big technology companies that dominate the AI field like Microsoft, Apple, Google, and Amazon.

Fair

In this thesis, a fair AI system is one that is beneficial for everyone and does not discriminate any groups such as gender, race, class, sexual orientation, disabilities, etc.

Gender

According to the Merriam-Webster dictionary, gender refers to socio-cultural factors such as “the behavioral, cultural, or psychological traits typically associated with one sex”. This is a differentiation from sex, which only consists of biological factors. (“Gender,” 2020) Although this thesis recognizes that gender can’t be separated from other issues of inequality such as race, disabilities, or class, this thesis focuses on the gender aspect due to the constraint of time and resources.

This thesis further recognizes that gender is not binary but a spectrum. Gender bias in this thesis mainly refers to bias against cis-women not cis-men, i.e. women who were born with female genitalia and who identify as women. This thesis does not intend to exclude transgender people from the issue of gender bias and recognize that transgender people also suffer from this issue along with cis-women. However, in order to limit the scope of the thesis, interview questions do not explicitly mention transgender people when asking about gender bias.

1.3 How the Thesis is Organized

This thesis reflects the author’s journey and realization of her own biases and assumptions. The concepts learned during the phase of writing the Reflection chapter are not retro-actively edited in the preceding chapters of this thesis (Chapters 2 - Background, 3 - Theory, 4 - Research Approach, 5 – Findings, 6 – Discussion). This is to gradually introduce any inexperienced reader to this topic and bring the reader on the same journey. Apart from the Introduction, this thesis aims to present the knowledge in the chronological order as it was discovered by the author.

2 Background & Related Work

This literature review provides an overview of the more than 200 articles assessed, of which about 150 are referenced in this thesis. AI is increasingly assisting our decision making. AI and its algorithms surround our everyday lives and affect our decision making from search engines, suggestions on dating apps, your credit rating, the price of your insurance or flight tickets (O'Neil, 2016). Algorithms also guide our decision for what to watch next on media platforms or which job ads we should see (Lambrecht & Tucker, 2019). Some data scientists would even go so far as to say that "algorithms decides who lives and who dies" (We All Count, n.d.-a).

With such great power vested to the AIs it is important that their conclusions can be trusted to be fair. However, as this literature review will show, the outcomes of AIs are at risk of being tainted with bias.

2.2.1 Literature Search

In order to find literature for this thesis the library search function, the Scopus database, and Google Scholar was used. The Snowball Method was also used to find relevant sources in the reference lists of relevant research articles. Some books and literature was recommended from my supervisor Maja Van der Velden, other peers, and people within the AI industry who knew about my thesis project. Additional literature was also found by researching other works of authors who had written about gender bias in AI.

Some literature was excluded because it was not peer-reviewed. A reference was included if it was continually referenced by several other sources. However, some sources that were not referenced by many others were still included because there may be other reasons as to why it was not referenced a lot. Reasons such as being newly published or because there are fewer doing research on this topic with a focus on gender, as opposed to research on general bias in AI. Additionally, the topic of gender bias in AI is somewhat new and the amount of research papers on this topic is limited compared to other areas.

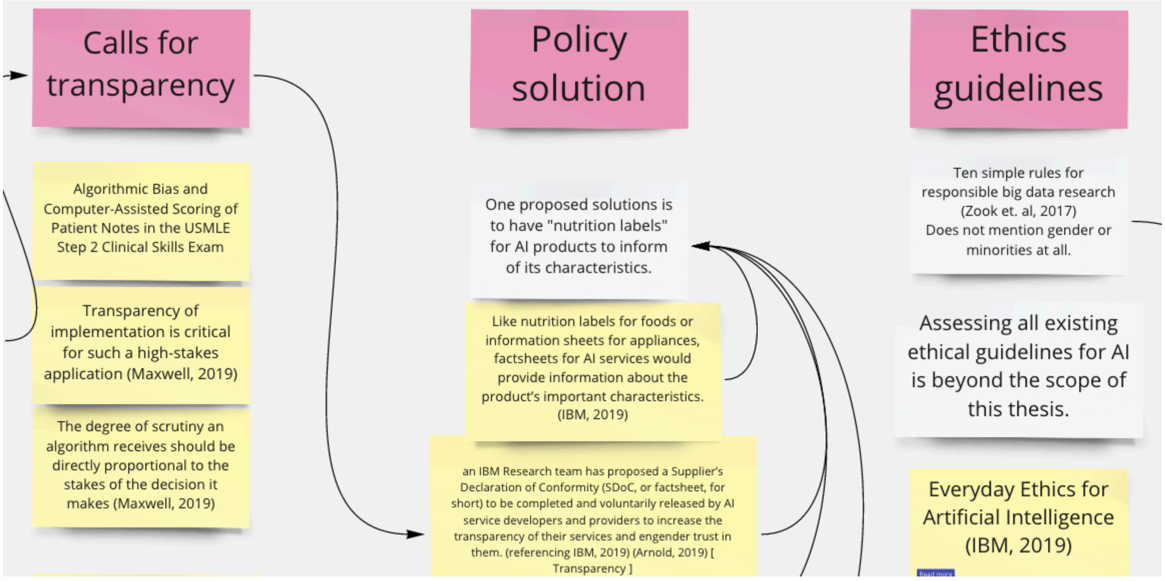
Scopus was mainly used because articles there are peer-reviewed and the database includes several other databases and publishers such as Elsevier, Springer, IEEE, ACM. Google Scholar was somewhat used but to a lesser degree since the University of Oslo has full-text access to Scopus articles.

The questions guiding the literature review were questions like why gender bias in AI is an issue, what is being done about it, why are some of the problems still an issue despite solutions being known, and does anyone use the published ethics guidelines. Not all these questions were answered by the literature review and gaps outlined in chapter 2.7 were identified.

The main keywords when searching for literature included: Gender bias in Artificial Intelligence, gender bias in AI, gender bias machine learning. More than 150 references are included in this thesis. Over 240 references and at *least* their abstracts have been read and saved to the Zotero reference manager.

Organizing Literature in Miro Board and Google Docs

Relevant reading notes and quotes for the literature review was first copy-pasted into Google Docs. The virtual board Miro.com was then used for organizing the reading notes according to related themes.



Calls for transparency

Algorithmic Bias and Computer-Assisted Scoring of Patient Notes in the USMLE Step 2 Clinical Skills Exam

Transparency of implementation is critical for such a high-stakes application (Maxwell, 2019)

The degree of scrutiny an algorithm receives should be directly proportional to the stakes of the decision it makes (Maxwell, 2019)

Policy solution

One proposed solutions is to have "nutrition labels" for AI products to inform of its characteristics.

Like nutrition labels for foods or information sheets for appliances, factsheets for AI services would provide information about the product's important characteristics. (IBM, 2019)

an IBM Research team has proposed a Supplier's Declaration of Conformity (SDoC, or factsheet, for short) to be completed and voluntarily released by AI service developers and providers to increase the transparency of their services and engender trust in them. (referencing IBM, 2019) (Arnold, 2019) [Transparency]

Ethics guidelines

Ten simple rules for responsible big data research (Zook et. al, 2017)
Does not mention gender or minorities at all.

Assessing all existing ethical guidelines for AI is beyond the scope of this thesis.

Everyday Ethics for Artificial Intelligence (IBM, 2019)

Figure 1. A screenshot showing an illustrative excerpt of the Miro virtual board that was created for organizing reading notes. See more details in this link: https://miro.com/app/board/o9J_ktfvCzk=/ (password: genderbias)

2.1 Discrimination and Bias in Artificial Intelligence

There are numerous cases and examples of biased AI algorithms (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016; Caliskan, Bryson, & Narayanan, 2017; Deshpande, Pan, & Foulds, 2020; Lambrecht & Tucker, 2019; Tatman, 2017; Zhao, Wang, Yatskar, Ordonez, & Chang, 2017). One of the more notable example is Amazon’s internal resume filtering AI that were to aid human resources in filtering job applicants (Dastin, 2018). The AI was trained on the resumes of the hired employees from the last 10 years. Dastin reports that the system would sort out all female applicants due to the history of Amazon hiring white men. It would sort out resumes that for instance had the line “President of Women’s Chess Team” or whether the resumes contained “active” words like “execute”. Such active words would be found more often on male resumes than female, due to the different writing styles of men and women.

Another notable example is the AI that would predict the likelihood of recidivism in prisoners. This AI would predict recidivism in African-American prisoners at higher rates than white prisoners because the AI system was trained on historical data (Angwin et al., 2016). The data was based on previous decisions that human judges had made. The analysis done by the news organization ProPublica showed that this data and AI system was biased against prisoners who were people of color (Angwin et al., 2016).

Multiple research articles and literature agree that AIs can contain bias and embed stereotypes (Bolukbasi et al., 2016; D’Ignazio & Klein, 2020; Leavy, 2018; Noble, 2018; S. M. West et al., 2019). Different types of discrimination such as sexism and racism are being built in to the AI systems (Crawford, 2016). Some even go so far as to say that AI is undoing women’s rights (Dejan Jotanovic, 2018; Leavy, 2018). More examples of biases in AIs are found in chapter 2.2.

2.1.1 AIs Are Not Objective

When removing the human from the equation of making decisions, one might assume that the human bias is no longer a problem. However, this is not correct, as AIs can inherit the bias of the humans who make them (M. West, Kraut, & Chew, 2019). According to a report by UNESCO and EQUALS, the unrecognized bias that is built into AI algorithms perpetuates and exacerbates gender inequalities (M. West et al., 2019). Faulkner believes that technology is gendered because those who make them are usually men (Faulkner, 2001). Although warnings about biases in Computer Systems date back to 1992, the realization that computerized systems are not objective just because they are machines, has not been more widely recognized until discriminatory systems were highlighted by researchers and came into the limelight of media (Bates, Clough, Jäschke, & Otterbacher, 2018).

As mentioned above, many AIs have shown to not be free of bias. Looking at AI through the lens of Winner's "Do Artifacts Have Politics?", it suggests that AI is not objective, and it might be an artifact with politics (Winner, 1980). Winner's article showed that the low-hanging bridges of Long Island that were designed so low that buses could not pass under them, although seemingly neutral, their engineering was in fact political (1980). According to Winner, The bridges prevented buses from reaching the beach, in practice preventing black people from going to the beach. Similarly, AIs can be racist or sexist and other forms of excluding and oppressive when it prevents groups of people from accessing and utilizing technology equally. This is demonstrated in the examples in chapter 2.2, Different Types of Bias in AI.

2.1.2 Biased Data Leads to Biased AI

There is inherent bias in AI because a lot of data is already biased (Zou & Schiebinger, 2018). There are many ways bias can enter the data. Biased data is one of the key problems of biased algorithms; AIs are only as good as the data you put in (Avila, Brandusescu, Freuler, & Thakur, 2018). This is supported by professor Iyad Rahwan of MIT: "Data matters more than the algorithm" according to Rahwan (Wakefield, 2018). He was a part of the team that developed Norman, dubbed *the psychopathic AI*.

Norman was the result of an AI project where the algorithm was trained on graphic images of violence and demonstrates the effect of data on the same algorithm. Norman was then compared to a standard image captioning AI system when shown Rorschach ink blots. The same ink blot was captioned by the other AI as "A group of birds sitting on top of a tree branch", whereas Norman saw "A man is electrocuted and catches to death." The comparison showed that despite being similar algorithms the "associations" of the image captioning relied heavily on whether the training data was regular cuddly kittens or from the "dark corners of the internet" (MIT Media Lab, 2018).

The organization We All Count outlines how data bias comes about. The entry point of data bias can begin at the stage of data collection. The bias begins with the funding of what kind of data should be collected (Abdalla & Abdalla, 2020; Wachter, Mittelstadt, & Russell, 2021; We All Count, n.d.-b). The funding also impacts the scope and scale of the data collection. If the sample is too small or the sample selection is skewed, the data is not likely to be representative (We All Count, n.d.-b; Zou & Schiebinger, 2018).

The factors then go on to who is collecting the data and how they behave while doing it (Robson, 2002; We All Count, n.d.-b). Another factor that is very relevant is the project design and methodology of the data collection (Robson, 2002; We All Count, n.d.-b). Data can for instance automatically be extracted from platforms and historical data, they can be collected via surveys or via manual interviews. For data collection via human-to-human methods such as surveys or interviews, cultural translation can play a role as to whether the data in question is captured (We All Count, n.d.-b). Regardless of cultural barriers, the definition of something that is being counted can change the result on the same data (Krause, 2019; We All Count, n.d.-b).

After the data has been collected bias can be introduced if the data has been corrupted (We All Count, n.d.-b). Further bias can be introduced in the analysis of the data and how the results are interpreted (We All Count, n.d.-b; S. M. West et al., 2019; Zook et al., 2017). When the data is being shared, bias can be introduced depending on what data is shared; omitted data can lead to bias (S. M. West et al., 2019).

The metadata for the dataset also impacts the bias of an algorithm (Zou & Schiebinger, 2018). If the data is not disaggregated and labeled, it can be difficult for developers and researchers to identify the gaps in the data (Perez, 2019). Zou and Schiebinger (2018) state that datasets should come with a data biography and labels because then developers and researchers can use that to find the source of bias. They suggest labeling the demographic information such as where the data or people in the data is from, gender, and ethnicity. Several open source datasets for voice recordings are not labeled with its demography, which might make it difficult for developers to assess the gender balance or what data might be missing (Perez, 2019). Additionally, missing data can lead to bias if the total amount of data is too little (O'Neil, 2016). Test data are usually as biased as the dataset used for training because they are normally just subsets of the training data (Zou & Schiebinger, 2018).

Not only is the data itself biased, but there might also be missing data that needs to be taken into consideration (Wachter et al., 2021). When we use data to understand public needs, we run the risk of missing out on the needs of those who are not within the data. The dataset can be biased because of missing data leading to it not being representative (Zou & Schiebinger, 2018). There could also be missing data in terms of lack of data on a certain demography. This can lead to an exclusion of their voices and needs in the system (Perez, 2019).

According to O'Neil (2016), data becomes more biased with a feedback loop tainted with bias. A biased decision feeds new data into what becomes a toxic feedback loop amplifying the bias in the AI system (O'Neil, 2016). The lack of a validation of the decisions that are made feeds new biased data back into the AI (O'Neil, 2016). See the next section for more about feedback loops.

2.1.3 Toxic Feedback Loops Amplify Existing Inequalities

There is common consensus among several researchers and data scientists that the decision making of AI often exacerbates existing inequalities (Benjamin, 2019; Eubanks, 2018; Noble, 2018; O'Neil, 2016; Stumpf et al., 2020; M. West et al., 2019). The imbalanced power structures that exist in AI further exacerbates inequalities in the rest of the world (Parsheera, 2018). Noble has coined the term *technological redlining* as "the ways digital decisions reinforce oppressive social relationships and enact new modes of racial profiling" (Noble, 2018, p. 1).

A biased decision will still be a part of the feedback loop that will amplify the effect of that bias (Zou & Schiebinger, 2018). This could lead to grave societal consequences on a big scale (O'Neil, 2016). Data scientist Safiya Noble (2018) agrees with O'Neil and has predicted that AI will become a major human rights issue.

Despite the aforementioned consensus among researchers that AIs often increase inequalities, a report sponsored by The Equality and Anti-Discrimination Ombud in Norway (LDO) shows that the Norwegian government encourages increased automation of decision processes. According to an Official Norwegian Report on a new law for governance they found that automated decision-making is especially advantageous for the marginalized people of society as they do not need to follow up on their applications and their claims for such things as benefits and student loans (NOU 2019: 5, p. 259). The government's view is alarmingly disturbing when most research in this literature review contradicts this.

2.1.4 The Reductionism of AI and its Limitations

In the race for innovative technology it appears that the impact of technology on humans has taken a backseat consideration while developing "the next big thing". IBM has stated in their AI report that "A tech-centric focus that solely revolves around improving the capabilities of an intelligent system doesn't sufficiently consider human needs." (IBM, 2019, p. 10). Such human needs could for instance be empathy, compassion, flexibility, and understanding. One of the benefits of human errors is that someone can be held accountable. However, the unclear distribution of accountability in AI can be utilized as a scapegoat for risky decisions that might be difficult to explain to the company leadership (Leicht-Deobald et al., 2019).

Another benefit of human decision-making is that although they are biased, there is at least a diversity of different biases (Benjamin, 2019). The danger of a widespread use of AIs for decision-making is that everyone using the same AI will inherit the same biases. If, for instance, all tech companies were to use Amazon's resume filtering AI, no tech company would hire any women (Dastin, 2018).

One of the weaknesses of AI is its deterministic nature where A always leads to B given that A is true (Leicht-Deobald et al., 2019). People are treated as numbers and all numbers are treated "equally" by the algorithm. It lacks moral imagination and the ability to compromise (Leicht-Deobald et al., 2019).

For instance, everyone applying to Lånekassen should get a loan and scholarship if they've gotten loans for less than 8 years and are no more than 60 credits behind, and those who don't fulfill those criteria should be rejected. However, if one of those conditions are not fulfilled due to for instance medical reasons, an exception can be made given sufficient documentation. But when one applies, one is directly rejected by the AI if both conditions are not fulfilled. When calling their customer service, you are once again asked to re-apply online. However, the AI rejects you once again because there is no way to communicate to the algorithm that documentation is on the way in the mail.

An AI cannot see the nuances and grey areas of ethics, and personal exceptions that a human could see are not taken into account. Here, a human might consider requesting additional documentation in the case where one of the requirements are not fulfilled to give the student an exception from these criteria. What is lost when humans aren't making the decisions? Leicht-Deobald et al. (2019) points out that when humans make mistakes it leads to learning experiences, but this might not be the case when the AI systems makes mistakes. There might be limitations to the "answers" algorithms provide us, however in the hype of AI, that seems to have been forgotten (We All Count, 2020).

Is AI the Best Solution?

One of the issues that have arisen in the AI revolution is the use of AI for seemingly incompatible purposes. Statistical decision-making and task performing can be good for narrow and clearly defined contexts, but can be damaging in cases where human judgement is required to reach good decisions. Human deliberation and judgement is required in gray area decisions that affect human lives (Leicht-Deobald et al., 2019). This could for instance be which student should be admitted to a university, what grade they should receive, or what candidate should be interviewed or hired for a position. An assessment of whether AI should be implemented at all for any given business use case needs to be made, and some AIs should not be made at all (S. M. West et al., 2019). Just because an organization *can* use AI, does not mean that it *should*. Not everything that can be done should be done (S. M. West et al., 2019).

2.2 Different types of biases in AI

Biased algorithms and AI can affect everyone but they particularly affect marginalized groups of society as researchers have found that biased AI systems mimic the discrimination that has existed throughout history (S. M. West et al., 2019). West et al report that AI systems are "systems of discrimination" as their task is to sort, organize, rank, and categorize (S. M. West et al., 2019, p. 6). However, they say, this discrimination is not equally distributed as they tend to hurt historically disadvantaged groups more.

Apart from how well-intended AIs can hold bias against certain groups, AI can also be used as a tool for malicious purposes. Such purposes span from targeting vulnerable and low-income people by showing them specific ads in order to unethically make money off of them (O'Neil, 2016), to creating fake news or pornography without the consent of the subjects by using *deep fakes* (Harwell, 2018). Deep fakes are videos that have been edited using AI to manipulate the video's original appearance. The face and voice of a person can be swapped to make it look like they have said or done something they have not. Although these are important issues related to bias and inequalities, they are outside the scope of this limited thesis and will not be discussed any further. Instead, four types of bias classified into the demographics it discriminates is outlined below: class bias, racial bias, gender bias, and LGBTQ+ bias.

2.2.1 Class Bias

The predictive policing algorithms that predict where the next crime is likely to happen send police units to neighborhoods with a history of higher crime rates. However, since they are sent to the neighborhoods where previous crime has been found, they are not necessarily sent where there is objectively more crime happening (D'Ignazio & Klein, 2018; O'Neil, 2016). These neighborhoods in the US tend to be the poorer neighborhoods of minorities, such as African-Americans ("PRE-CRIME," n.d.).

When the predictive policing AI sends more police units to minority neighborhoods, the police will register more crimes from that area, which then becomes new data for the AI (O'Neil, 2016; Richardson, Schultz, & Crawford, 2019). This in turn becomes a toxic feedback loop; for each new crime found in such neighborhoods, the bias against poor communities in the algorithm increases (O'Neil, 2016; Richardson et al., 2019). Meanwhile, there might be *White-collar crimes* occurring in other richer areas without the police investigating it because they are not present and the crimes might go unreported as they are less noticeable (D'Ignazio & Klein, 2018).

There were many newspaper articles written about how Ofqual's predictive algorithm for A-levels in England led to widespread protests amongst students in England (BBC, 2020). Students chanted "Fuck the Algorithm" to protest the downgrades that would affect their university admissions (BBC, 2020). Grades were partially determined based on the previous performance of the school a student attended. Private school students were then more likely to get A's and less likely to be downgraded compared to disadvantaged areas (Richard Adams & McIntyre, 2020). The algorithm was used as COVID-19 led to the cancellation of the regular A-level tests which led to one third of the students being downgraded (Naughton, 2020).

Furthermore, Ofqual's algorithm made sure that the grades given for 2020 were of the same distribution from the three previous years factoring in attributes such as gender and ethnicity (Harkness, 2020). I.e. colored female students in 2020 were to get similar grades as colored female student from the past. Depending on how Ofqual designed their algorithm, these factors could potentially compound and lead to worse grades for a student whose demographic background included several factors that the algorithm would relate to worse grades.

2.2.2 Racial bias

The predictive policing feeds into the racial bias that exists against people of color. As mentioned in 2.1, a striking example of racial bias in AIs is Northpointe's software system for risk assessments (Angwin et al., 2016). It is an AI system used by judges to get recommendations on whether a prisoner should be let out of prison and predicts the probability that a prisoner will reoffend. Angwin et al. found that such risk assessment systems have a tendency to recommend letting out white prisoners at a higher rate than black prisoners.

Other examples of racial bias in AI are the facial recognition and image classification AIs. There are several examples of libraries and software for facial tracking and facial recognition which are unable to detect dark skinned faces (Buolamwini, 2018; Crawford, 2016). Researcher Joy

Buolamwini was forced to put on a white mask in order for the AI to be able to track her face. The pictures of iconic women such as Oprah Winfrey or Michelle Obama are categorized as “men” or something other than a black woman (Buolamwini, 2018). A study by Buolamwini and Gebru (2018) found that image classification AIs have a higher error rate for people of color, and even higher for black women.

A study investigating speech recognition systems made by Microsoft, IBM, Amazon, Google, and Apple, found that they all were less likely to understand the black speakers compared to white speakers (Koenecke et al., 2020). The findings made by Koenecke et. al. indicate that there is a barrier for equal use of this technology.

The photos of a Black user was tagged as "gorilla" by Google Photos in 2015 (Noble, 2018), and there were also other examples where images of Asians were categorized by the AI as a photo of people with closed eyes (Crawford, 2016). When Apple launched facial recognition for unlocking their iPhones the feature did not work as well for Asians (Papenfuss, 2017). It was not able to tell Asians apart as well as it did for white users and there were instances where Asian siblings could unlock each other’s’ phones (Papenfuss, 2017).

Data scientist Safiya Noble (2018) found that Google searches for the terms "black girl" and "white girl" would result in very different results that reflect racial stereotypes (2018). The search engine algorithms would return porn sites when searching for “black girl” or “black women”, whereas the terms “white girl” or “white woman” did not lead to porn being the top results. According to Noble, these search results exacerbate the stereotypes and oppression of colored women being seen as subhuman objects. According to Benjamin, intentional harm is not needed for racism to be embedded in the tech industry, all it takes is to not be aware of the past and how it impacts the present (Benjamin, 2019).

2.2.3 Gender bias

The previously mentioned AI for shortlisting job applicants at Amazon (2.1) is one example of gender bias in AI (Dastin, 2018). There are numerous examples of AIs biased against women. Voice assistants such as Apple’s Siri and Microsoft’s Cortana have been a hot topic of debate. Most AIs were launched with only a female voice and some question as to how this will contribute to the stereotype of the quiet and obedient female secretary (Rachel Adams, 2020; Dejan Jotanovic, 2018). One report by UNESCO and EQUALS found that the design of voice assistants reinforces gender bias and the female stereotypes of women being subservient (Specia, 2019; M. West et al., 2019). The report found that if the AI were given the command "You're a slut", then Alexa would respond with "Well, thanks for the feedback", whereas Siri responded with "I'd blush if I could" (M. West et al., 2019, p. 107).

Studies have found that machine translation leads to gender biased results (Bolukbasi et al., 2016; Caliskan et al., 2017). A study by Caliskan et al. found that Google Translate changes gender-neutral pronouns in languages like Turkish, Finnish, Persian, and Hungarian to stereotypically gendered ones in common languages like English, German, French, Russian, and Spanish (2017). Google chooses the pronouns that appear most frequently with a word and translates gender-neutral sentences into "He is a doctor. She is a nurse." (Caliskan et al., 2017). Similarly, a study by Rachael Tatman (2017) found that YouTube's automatic captioning system were less likely to understand female speakers.

Another striking example of bias in AI is a study that found that an image classification AI would classify a bald man in the kitchen as a woman, because the trained association between women and kitchens were so closely linked (Zhao et al., 2017). A study found that image search results amplified gender stereotypes in occupations and underrepresented women somewhat (Kay, Matuszek, & Munson, 2015).Similarly, a study showed that word embeddings that were trained on Google news articles associated "man" with words like computer programmer, architect, philosopher, protégé, superstar; whereas "woman" was associated with occupations and terms like homemaker, nurse, receptionist, hairdresser, and diva (Bolukbasi et al., 2016).

Dynamic ad algorithms use demography and history of clicks to target their audience. A study found that such algorithms has led to job ads in STEM being shown less to women (Lambrecht & Tucker, 2019).

Although most examples of gender bias are bias against women, a study by Thelwall (2018) found gender bias against men. AIs are biased against whoever there is less data of, but also the nature of the data. Even when there were equal amounts of data from male and female reviewers, the writing style of women made their opinions more prevalent in a summary than men's (Thelwall, 2018).

2.2.4 LGBTQ+ bias

Uber's facial recognition system for drivers have more difficulty recognizing trans faces (Melendez, 2018) and AI that is used to predict sexuality could be abused in a homophobic "witch hunt” for homosexuals (Schei, 2020). The airport body scanners struggle with scanning trans people because the gender setting on the scanners are limited to “female” and “male” (Costanza-Chock, 2018). The result of this binary setting is that trans people are flagged because their body proportions in the chest or groin area might not comply with the expectations of the settings (Costanza-Chock, 2018).

2.3 Entry Points of Bias in AI

This section outlines the different ways general bias can enter AI according to the literature. These entry points also include the entry points of gender bias.

2.3.1 Developers Are Biased

The gender and racial disparity in AI affects the AI products developed. Some researchers refer to this issue as 'AI's White Guy Problem' (Crawford, 2016). Most people carry some unconscious bias, but some developers have explicit bias against women and even publicly declare these views (S. M. West et al., 2019). Noble questions how we are to believe that algorithms are neutral when their creators certainly are not (Noble, 2018).

Developers have a hand many of the entry points for bias. For instance, it is necessary for developers to question whether there is a real correlation between the available data and the output that the AI is looking for. In Amazon’s example there seems to be little to no correlation between resumes and the qualification of candidates as not only did it reject qualified candidates, it also recommended unqualified ones (Bubakr & Baber, 2020; Dastin, 2018).

Bias might be introduced in the cleaning of the data (Jones, 2018). Data cleaning consists in a lot of formatting and addressing gaps of missing data. It can also include omitting parts of data that are deemed irrelevant for the purposes of its use. For instance whether to include phone numbers, names, or gender from a resume. The feature selection for the training of the algorithm might introduce bias. Bias might be introduced if they don’t identify features that need to be actively ignored during training, such as gender in the Amazon example (Dastin, 2018).

During testing and validation they are the ones who set the success criteria for a passed test (S. M. West et al., 2019). As mentioned previously, test data is usually as biased as the dataset for training because they are both split from the same initial dataset (Zou & Schiebinger, 2018).

Additionally, developers need to design the the user interface and decrease the automation bias (Sharkey, 2014) or the lack of data literacy of the user (Leicht-Deobald et al., 2019).

2.3.2 The Choice of Algorithm and Fairness Metrics

The choice of mathematical algorithms and machine learning models affect how data is processed, and can therefore affect whether potential bias is amplified or mitigated (Zou & Schiebinger, 2018). According to Zou and Schiebinger (2018), a standard machine learning algorithm will optimize for any individuals that are more often represented in the training dataset because this will raise the performance.

Let's look at closer at the example with the Amazon AI for filtering resumes. Dastin (2018) reports that when training the model it can find patterns between demography and the hired employees and mistake that for a success criteria. For instance, it found that resumes that are male or go to colleges that are not women's colleges tended to be hired at a higher frequency (Dastin, 2018). It incorporates therefore this bias in the star ratings that HR used to sort the applicants (Dastin, 2018).

Researchers at Google and Stanford found that using a GAN was effective for mitigating bias against a demography (Zhang, Lemoine, & Mitchell, 2018). Like a regular model, the network takes an input and produces a prediction, in the example of Amazon, the equivalent would be the candidate star rating. But it also has an adversary that simultaneously tries to model a variable, like gender. The objective during training is for the model's ability to predict the star rating to increase, while the adversary's ability to predict gender to decrease. The goal for training is for the adversary to not be able to distinguish a male candidate from a female candidate. If the network can predict star ratings without being able to distinguish the gender of the candidate, then this will mitigate existing gender bias.

Furthermore, *overfitting* on biased data might create a more biased machine learning model (Kakarmath et al., 2020; O'Neil, 2016). When training a machine learning model, one has to decide when to stop training by defining the optimal error rate. If the error rate is zero, it means that the model is an exact replica of the data. The goal is to decrease the error rate, but if the error rate is too low the risk is that it picks up noise in the data that look like patterns, but does not actually depict reality (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014).

An example of such noise would be how the Amazon's resume filtering AI found a pattern between being a woman and not being qualified for their job openings. An overfitted algorithm would then only be able to correctly predict on the data it was trained on and not be able to reach accurate predictions on unfamiliar test data or new data from the real world (Schaffer, 1993). This issue is referred to as overfitting. Overfitting on biased data might increase the bias because it replicates the bias in the data (Kakarmath et al., 2020; O'Neil, 2016).

A recent paper by Wachter, Mittelstadt & Russell (2021) from Oxford University defined a way to assess whether classifiers for fairness metrics preserved the status quo of the data or would take into account historical inequalities and balance for this. These terms were coined as *preserving bias* and *transforming bias*, respectively. They argue that choosing a fairness metric that would maintain the status quo that is in the data is *not* neutral. They further say that practitioners should assume that the data and system is biased, and the only exception is where thorough testing have been done and it is certain that the area of application has never had a history of inequalities. Their paper shows that even if one were to assess bias in an AI system, which fairness metrics that are chosen for measuring the bias in the algorithm can also affect the level of bias introduced in a system (Wachter et al., 2021).

2.3.3 End users are biased

End users are affected by human bias when they interpret the data presented by the AI. End users can have conscious or unconscious bias against a certain demography or they can be affected by different cognitive biases. Three studies suggest that people have a *bias blind spot* where they think they are less prone to bias than others (Pronin, Lin, & Ross, 2002). Users might be affected by *in-group bias* which makes them see people who are similar to them as better candidates, which in turn might affect their hiring practices (Huston, 2018). Biased hiring practices could lead to biased data on who is an ideal candidate.

However, the biggest issue when it comes to cognitive biases in using AI systems is automation bias. Automation bias is when the user depends too heavily on the system to make the correct judgement and assume that the system's recommendation is correct (Sharkey, 2014). This bias leads to users to not look for opposing information and to not use their own judgement to validate the conclusion of the system.

The way the system is designed and what kind of data is presented can enable or discourage automation bias. How the data is visualized and presented affects the interpretation of the end user

(Dodge, Liao, Zhang, Bellamy, & Dugan, 2019; We All Count, n.d.-b). Furthermore, the end user's ability to make a good decision based on the data can also depend on their data literacy and their awareness of potential bias (Leicht-Deobald et al., 2019).

The aforementioned human biases might affect the decisions of the users. O'Neil (2016) explains how these decisions become the data that is fed back into the system. She states, if the decisions are biased, then this becomes new biased data that is fed back into the AI which makes the AI more biased. As previously explained, the cycle of biased decisions becoming new data and being fed back into the AI system leads to a toxic feedback loop for bias, which is what O'Neil (2016) refers to as Weapons of Math Destruction. The entry points outlined in the previous chapters amplify each other as they contribute to a biased decision that becomes a part of the toxic feedback loop (O'Neil, 2016).

The problem is not only that a toxic feedback loop leads to an increase in algorithmic sexism. Results that are presented to users and confirms their existing biases also increases *their* gender bias. A study demonstrated that gender stereotypes in search engine results both confirmed and exacerbated existing gender stereotypes of participants (Kay et al., 2015).

2.3.4 Power Structures That Enable Bias

The research of Noble, Raji and Boulamwini (2018; 2019) shed an unflattering light on the AI giants which pressured them to make changes. The tech giants *did* make changes to their algorithms, however, Raji and Boulamwini (2019) question whether these changes were applicable only to the dataset they published Noble (2018) questions what processes go on behind the scenes to push change and what would have happened whether no one finds out about the problems.

The issue of gender bias in AI lies not only in the data or the algorithms of AI, but also within the surrounding power structures that enables such biases to remain (West et al., 2019). Researchers argue that data biases reflect and are affected by the power imbalances and institutional racism and sexism that are present in the AI companies and government institutions (Lazovich, 2020; Zou & Schiebinger, 2018). The majority of positions in AI and CEOs in AI are held by men who are likely less affected by this issue, and this gender disparity might be a contributing factor as to why it is not being addressed (United Nations University & EQUALS, 2019; S. M. West et al., 2019).

Google fired ethical AI team leads after censoring incriminating research (Johnson, 2021; Jonhson, 2020). Timnit Gebru is a former Google-employee who was fired in relation to the research she was doing on the negative social consequences of large language models in AI (Jonhson, 2020). Consequences such as perpetuating racism, carbon footprints, and increased costs and entry barriers for deep learning research . After Gebru was fired, the ethics team lead Margaret Mitchell was fired because she openly criticized Google for firing Gebru (Johnson, 2021). A study found similarities between the current AI industry and how the tobacco industry funded academic research to impact the research agenda and to lobby their interests (Abdalla & Abdalla, 2020).

2.4 Proposals for Solving General Bias in AI

2.4.1 Ethics Guidelines

This chapter will give a brief introduction to the landscape of ethical guidelines for AI but will not present any ethical guideline in detail as such an analysis is outside the scope of this thesis. There are indications that ethics guidelines are important and necessary for addressing the ethical issues of AI (Jobin, Ienca, & Vayena, 2019; Parsheera, 2018).

Many guidelines for ethics in AI have been published (Gordon-Murnane, 2018; IBM, 2019; Jobin et al., 2019; Zook et al., 2017). A study from 2019 found 84 documents outlining ethical guidelines for AI, most of them released in the US or in the EU (Jobin et al., 2019). Perhaps one notable guideline is the *Asimolar AI Principles* that in 2020 so far had been signed by 1677 researchers and 3662 other interested parties in AI, including the late Stephen Hawking and Elon Musk (The Future of Life Institute, 2017). Many big AI actors such as IBM, Microsoft, Google-owned DeepMind, the Internet Society, World Economic Forum, UNESCO, and similar organizations have published ethics guidelines (Jobin et al., 2019).

At least one of the ethics guidelines explored while conducting this literature review appear to be high-level and without concrete solutions or practices to avoid unethical AI systems (The Future of Life Institute, 2017). High-level guidelines leave the interpretation of the guideline to the creator and as such, the ethical outcome depends on the personal judgements of the creator. Parsheera (2018) writes that because the concept of fairness can be interpreted in multiple ways, it might lead to AIs that are less fair. One of the challenges seem to be that ethics guidelines do not easily translated to practical processes. Parsheera (2018) suggests investing in tools that can translate ethical guidelines into concrete practices. AI systems have been developed for the purpose of an automated testing for bias (IBM Developer Staff, 2020).

In their thematic analysis of the landscape of AI ethics guidelines, Jobin et al. (2019) found that the guidelines had big differences in several aspects, such as how they are understood, why they are important, and how they should be implemented. They also found that sometimes the principles would contradict each other, and they say that more information is needed on which principles to prioritize or how to deal with such contradictions (2019).

Regardless of how many AI ethics guidelines there might exist, it is not clear whether they are actively being used. Within this body of literature it was not possible to find any reports with information on whether they are adopted or used by AI practitioners.

2.4.2 Policy solutions for transparency

Calls for accountability and transparency

AI companies appear to make money regardless of whether their AI systems are fair or discriminatory. AI systems are opaque and are not audited for fairness or equality by external independent parties. One might think that life-altering decision making processes ought to be overseen by a responsible authority. However, with a lot of systems there is no way to investigate whether the decision the AI proposed is the best one. Most AI systems are considered proprietary software and can avoid the scrutiny of the public under the rights to not disclose the recipe of the “secret sauce” (Thelwall, 2018).

The discovery of an AI’s discriminatory practices is then perhaps left to curious researchers (Noble, 2018; O’Neil, 2016), investigative journalism (Angwin et al., 2016), anonymous whistleblowers (Dastin, 2018), or chance like when a MIT student could not test her facial recognition project without a white mask (Buolamwini, 2016), and public perception in the case of a Chinese woman unlocking her colleague’s iPhone using Face ID (Papenfuss, 2017).

Even if it is discovered that an AI system is discriminating Norwegians, the discriminating system is not required to change because like Big Tech companies, most AI companies are outside of Norwegian jurisdiction. AI programs are considered proprietary software and are not transparent. The leaders in AI are ad companies like Google and Facebook and they are profit focused. Despite Google’s slogan being “do no evil” their search engines do contribute to societal harm (Noble, 2018). One proposed solutions is to have "nutrition labels" for AI products to inform of its characteristics (Arnold et al., 2019). Zou and Schiebinger (2018) suggest that conference organizers should require metadata as a part of the paper submission process.

In May 2020, the Norwegian government announced a new regulatory sandbox environment within the Data Protection Authority for the testing and development of AI systems (Datatilsynet, 2020). The Data Protection Authority (2020) reports that the initiative will be the second AI sandbox that is within a Data Protection Authority in the world following Great Britain.

2.4.3 Development and Design Solutions

Explainable AI

AI is a black box and we don't know how it makes its decisions (Timcke, 2020). Efforts are being made to find ways of explaining the black box of AI. However, some criticize and say that the transparency created by these efforts so far have not been usable or practical in a way that would be helpful for people (Abdul et al., 2018).

Research suggests that different explanation styles impact how people judge the fairness of an AI system (Dodge et al., 2019). This study found that depending on how the decisions of an AI are explained and depending on the beliefs of the subjects, they would judge the fairness of the AI differently. These findings might mean that even if the processes of an AI are explained, an AI might only become more fair in the eyes of some. This demonstrates the difficulty of creating a universally fair AI system even if the AI is explainable since people's definitions and perceptions of fairness vary.

Including End Users in the Design Process

Fixing faulty AI systems after they have been implemented and put to use is likely more difficult and complex than to fix issues at the design stage. IBM's ethical guidelines for AI cite Frank Lloyd Wright in this quote: "You can use an eraser on the drafting table or a sledgehammer on the construction site" (IBM, 2019, p. 8).

The challenge with this, is that the outcomes of technology can sometimes be difficult to predict. The Design Justice Network Principles (2018) and Costanza-Chock (2018) suggests centering the voices of those who are impacted by the resulting AI systems. Leicht-Deobald et al. (2019) advise companies to include employees in the process of acquiring an AI system *before* it is acquired. This way, they say, important considerations that the technical department are not aware of can be brought to light before it is too late to change the system.

AI needs to be fixed first and used later, and before it becomes a human rights issue (Noble, 2018). The iHuman documentary on AI and ethics questions what will happen if we do not intervene now (Schei, 2020). At the current pace of how fast AI evolves, iHuman questions whether there is a point of no return where democracy is eroded and the power to change will lie in the powerful hands of the few. An example illustrated in iHuman is that a dictator with autonomous killing drones with facial recognition for target identification would have a dangerously efficient military power.

Data2x (2020), a non-profit that works for more equity in data argue that it is better to deliberately choose features, rather than just feeding data into a neural network that chooses the features based on the data. Additionally, they recommend slowing down to assess the processes for equity (Data2x, 2020).

An AI that checks for Fairness

One of the technical solutions for solving bias in AI are open source toolkits for fairness (Bellamy et al., 2018). One example is the IBM AI Fairness 360 Toolkit software (IBM Developer Staff, 2020). The IBM 360 Toolkit offers a repository of algorithms that supposedly can detect and mitigate bias in other AIs. IBM's toolkit are based on 11 research articles on bias mitigation and contain 9 different mitigation algorithms (Bellamy et al., 2018). IBM state that they aim to bridge the gap between the practice of AI developers and the knowledge of AI researchers (IBM Developer Staff, 2020).

The aforementioned study by Dodge et al. (2019) showed that the perceived fairness changed depending on whether they were shown the features that contributed to that conclusion, or whether they were explained the process of the algorithm. The participants' judgements of the fairness also changed depending on whether they were presented different outcomes when features such as race was changed.

The study conducted by Dodge et al. (2019) shows that the perception of fairness is subjective and fluid, which makes the notion of a Fairness Toolkit faulty. The fluidity of fairness makes it difficult, if not impossible, to code. It requires a succinct definition but it is not clear whether a such definition should be determined by companies such as IBM. The Fairness Toolkit is limited by the definition of fairness of the programmer or perhaps the person who pays the programmer. The programmer decides the features of fairness and creates the tests of when fairness has been fulfilled. This body of literature could not find any data on how widespread the use of such toolkits are.

2.5 Gender-Specific Proposals for Solving Bias in AI

2.5.1 Technical Solutions

There have been multiple scientific papers published on how to mitigate bias in AI systems using other algorithms (Celis, Huang, Keswani, & Vishnoi, 2019; Kearns, Neel, Roth, & Wu, 2018; Pleiss, Raghavan, Wu, Kleinberg, & Weinberger, 2017; Zemel, Wu, Swersky, Pitassi, & Dwork, 2013; Zhang et al., 2018). Some scientific papers have also been published using algorithms to specifically combat gender bias or debias algorithms (Deshpande et al., 2020; Jegadeesan, 2020). Even if many of the solutions cited in this paragraph have been combined into the IBM AIF360, there seem to be no universal solution to the issue of bias and gender bias in AI.

Leavy (2018) criticizes that technical solutions are not sufficient for AI systems that have been trained on text. She argues, a simple algorithmic fix will not take into account the historical gender ideology that is embedded in the language. She believes that developers need to be aware of the inherent gendering of words, such as policeman or housewife, in order to mitigate the bias that exists within the language. Algorithmic audits conducted by independent entities have been suggested as a solution; algorithms are then tested against benchmarks that are more balanced on both race and gender (Raji & Buolamwini, 2019).

2.5.2 Increase diversity

Programming used to be considered a women’s profession in the 1950’s because women were considered be more patient than men, and better at being detail-oriented (Kristiansen, 2020; Thompson, 2019). By 1984, 40% of the computer science students in the US were female (Kristiansen, 2020; Thompson, 2019). However, according to Thompson, 1984 was the year when women were subsequently pushed out of programming (2019). The field of ICT has suffered from a lack of diversity ever since. The number of female ICT professionals are estimated to be 16% in Europe, 22% in the Americas, and 26% in Asia in 2016 (United Nations University & EQUALS, 2019).

There are different estimations of the number of women in AI, but some claim that the disparity is even larger in the field of AI (S. M. West et al., 2019). Weissman estimates that the whole field of AI only had 13.5% women in 2016 (as cited in United Nations University & EQUALS, 2019, p. 96). At Google, 21% of the technical employees are female, but only 10% of the employees who work on machine intelligence are women (M. West et al., 2019, p. 19). It is estimated that only 12% of the leading machine learning researchers are female based on the number of attendees at the most prominent machine learning conferences in the world (M. West et al., 2019, p. 19). Whereas the World Economic Forum reports the number of female AI Professionals in the world to be 26% ("Global Gender Gap Report 2020," 2019). 15% women is the level of which *critical mass* is defined to accelerate change and improve conditions in science (Etzkowitz, Kemelgor, Neuschatz, Uzzi, & Alonzo, 1994). However, the level of gender balance needed for women to have a significant influence, improve team performance, and to be more than tokens or symbolic representatives of women is at 35% (Schwartz-Ziv, 2017).

Increased diversity in AI is proposed to mitigate the issue of gender bias (Avila et al., 2018). Several researchers suggest that the gender disparity in AI is contributing to the gender bias in AI and that the impact might even reverse the progress of equality (Leavy, 2018; S. M. West et al., 2019).

2.5.3 Awareness and Activism

In addition to academic research there is also activism to create awareness around the issue of bias and gender bias. Since around 2015 several books have been released to increase the awareness of the ethical issues of algorithms and AI. Among them are at least one Norwegian book, which is about digital ethics for AI by philosophers Bergsjø & Bergsjø (2019).

In English there are several books that are frequently cited in the literature. There is *Automating inequality: how high-tech tools profile, police, and punish the poor*, which is about the use of AI for predictive policing and how that increases inequality for the underserved communities (Eubanks, 2018). Data scientist Cathy O'Neil's *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* goes through several examples of how the use of big data algorithms can lead to less fair outcomes than previous traditional methods (2016). O'Neil's breadth of examples contain a deep technical understanding and is often cited by other papers on these issues.

The age of surveillance capitalism: the fight for a human future at the new frontier of power (Zuboff, 2019) is about how algorithms map and predict our behaviors and how these "futures" are sold to the highest bidder. *The black box society: the secret algorithms that control money and information* (Pasquale, 2015) and the aforementioned book by Shoshana Zuboff aim to shed light on how the big tech companies make massive amounts of money, gain immense power, and how these phenomena negatively affect our society. Pasquale's book seems to be one of the first books to be found in this body of literature on the topic of unethical algorithms.

Several books have also been released to increase awareness around gender bias in AI specifically. *Algorithms of oppression: how search engines reinforce racism* written by data scientist Safiya Noble (2018) and *Race after Technology: Abolitionist Tools for the New Jim Code* by sociologist Ruha Benjamin (2019) are both about how algorithms can reinforce racism and deepen inequalities. However, Noble's book writes more about how algorithms reinforce stereotypes of black women specifically. *Invisible Women: Exposing Data Bias in a World Designed for Men* written by journalist and activist Caroline Perez lists many examples of how algorithms and AI discriminate women and its negative effects (2019). Perez' book won the 2019 *Business Book of the Year Award*, awarded by the Financial Times and McKinsey (McKinsey & Company, 2019).

Documentaries about the issue of bias in AI has also been made to bring the issue to light (Kantayya, 2020; Schei, 2020). Norwegian director Tonje Schei interviewed several notable figures in the world's AI community and made the documentary *iHuman*. Different ethical aspects of the new AI world are highlighted in the movie such as the power autonomous killer drones can unleash in the hands of a dictator, the missing right to be deleted when there is an AI that scrapes social media platforms and websites for information and photos in order to create profiles of all human beings, and the research project that trained an AI to recognize which sexual orientation a person in a photo allegedly had.

The documentary Coded Bias (Kantayya, 2020) follows the journey of MIT researcher Joy Buolamwini from her first experience with racist facial recognition technology, to the preceding events, such as founding the non-profit organization Algorithmic Justice League ("Spotlight - Coded Bias Documentary," n.d.). Among those profiled are previously mentioned authors Cathy O'Neil, Virginia Eubanks, and Safiya Noble. Another notable profile in the cast is researcher Timnit Gebru who was fired from Google for her research on the social consequences of large language models, as mentioned in the earlier chapter 2.3.4 Power Structures That Enable Bias. The movie became available on Netflix in April 2021 (*Coded Bias - Global Release Marketing Toolkit*, n.d.).

Several non-profit organizations aim to spread information and work to advocate and address the issue of gender bias in AI systems (Algorithmic Justice League, 2020; We All Count, n.d.-a). We All Count is a project for equity in data science, they are trying to educate people working with data science, and have tools and training on how to avoid bias _(We All Count, n.d.-a).

Algorithmic Justice League has several initiatives, such as their involvement in the making of Coded Bias.

An example of the lack of awareness in the Norwegian AI field is to be found in a presentation of the project of collecting data samples of various Norwegian dialects (*Språkbehandling og kunstig intelligens*, 2020). This project is conducted by the Norwegian National Library. To their credit, it was mentioned in this presentation how the data they collect was labeled and gender dis-aggregated. However, when asked *why* they did this and *why* the differentiation in gender and age mattered, the person representing this project did not know and were not able to explain why, other than that this practice was "normal" (*Språkbehandling og kunstig intelligens*, 2020, pt. 55:39). This presentation is an example of the lack of awareness in the field, even when they do things right.

2.5.4 Interdisciplinary Collaboration

There might be a lack of knowledge about social sciences such as psychology, society, gender studies, or ethics amongst technical AI practitioners because they are more likely to have a background in engineering. Several researchers argue that progress will only be made in collaboration with experts from fields such as social studies, gender studies, psychology, humanities, and law (Abdul et al., 2018; Zou & Schiebinger, 2018).

Leavy (2018) criticizes studies that aim to reduce gender bias in AI systems without considering the research from the last few decades on how gender ideology is embedded in language. She states, involving other disciplines in the field of AI could help balance this knowledge gap. To avoid unconscious biases, Crawford (2013)suggests learning from the discipline of social science which has a tradition of questioning which biases they bring into a study. She suggests combining big data studies with the methods of smaller qualitative studies.

2.6 Model of Entry Points of Gender Bias

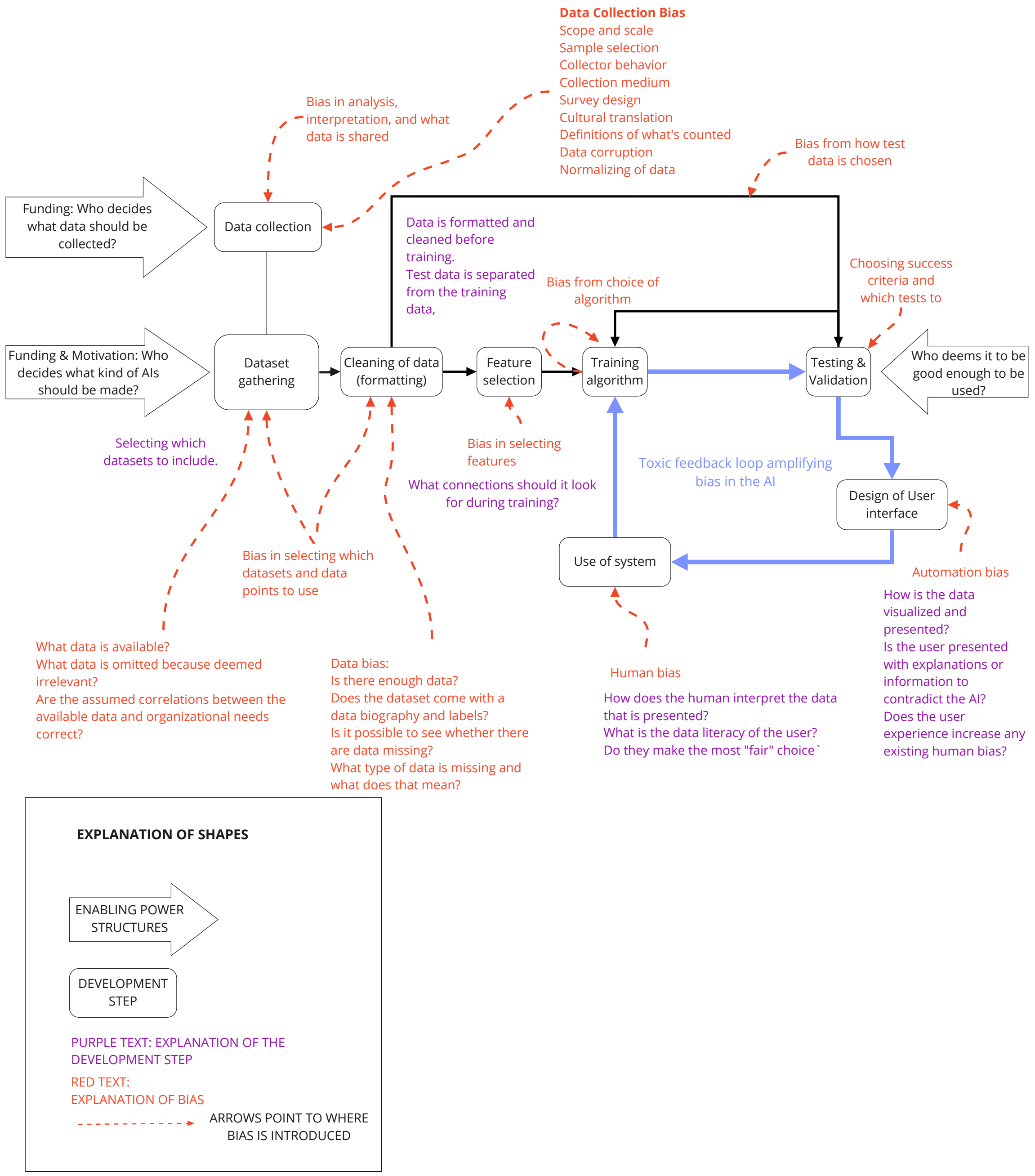

Data Collection Bias
Scope and scale
Sample selection
Collector behavior
Collection medium
Survey design
Cultural translation
Definitions of what's counted
Data corruption
Normalizing of data
Bias in analysis, interpretation, and what data is shared
Bias from how test data is chosen
Funding: Who decides what data should be collected?
Data collection
Data is formatted and cleaned before training.
Test data is separated from the training data,
Bias from choice of algorithm
Choosing success criteria and which tests to
Funding & Motivation: Who decides what kind of AIs should be made?
Dataset gathering
Cleaning of data (formatting)
Feature selection
Training algorithm
Testing & Validation
Who deems it to be good enough to be used?
Selecting which datasets to include.
Bias in selecting features
What connections should it look for during training?
Toxic feedback loop amplifying bias in the AI
Design of User interface
Bias in selecting which datasets and data points to use
Use of system
Automation bias
How is the data visualized and presented?
Is the user presented with explanations or information to contradict the AI?
Does the user experience increase any existing human bias?
What data is available?
What data is omitted because deemed irrelevant?
Are the assumed correlations between the available data and organizational needs correct?
Data bias:
Is there enough data?
Does the dataset come with a data biography and labels?
Is it possible to see whether there are data missing?
What type of data is missing and what does that mean?
Human bias
How does the human interpret the data that is presented?
What is the data literacy of the user?
Do they make the most "fair" choice`
EXPLANATION OF SHAPES
ENABLING POWER STRUCTURES
DEVELOPMENT STEP
PURPLE TEXT: EXPLANATION OF THE DEVELOPMENT STEP
RED TEXT: EXPLANATION OF BIAS
ARROWS POINT TO WHERE BIAS IS INTRODUCED

Figure 2. Entry points of biases in AI systems. The figure is mainly based on the literature review, and supplemented using figures from unpublished research from a conference presentation by Giannoumis & Bui (2019). The reader is advised to see the more comprehensive version with references in Appendix B (suitable for printing) or online for highest resolution: https://miro.com/app/board/o9J_ktfvCzk= (password: genderbias).

2.7 Gaps in the Literature

It is apparent from the literature that there is an issue of gender bias in AI. Below are some of the identified knowledge gaps that needs to be filled in order to move the progress forward.

The literature suggests that there is a lack of an interdisciplinary approach and knowledge to the problem of gender bias in AI. Leavy's article about the need for gender theory in solving gender bias in AI systems trained on text highlights the deeper issues of the inherent bias that exist in the language. Knowledge that probably only a feminist or linguist would know.

The knowledge gap created by this lack of interdisciplinarity is for instance the lack of an intersectional view when solving AI systems for gender bias (Costanza-Chock, 2018). The examples on the demo of IBM 360 Fairness toolkit separate race from gender and does not appear to have an intersectional view to assess fairness. The problem of racism and sexism is so intertwined that it is difficult to solve one without addressing the other. We need to know more about how we can use an intersectional view to solve the issue of gender bias.

Many ethics guidelines have been published in an attempt to address the issue of gender bias amongst other things. However, we know little about whether the guidelines are being used and the practices of AI practitioners. How does gender bias occur? Why does gender bias occur? Why is it not fixed? Is the issue actual ignorance or willful ignorance? Is the issue that the solutions do not exist or is it that they are not being used? Is the problem that AI practitioners do not know how to put ethical guidelines into practice, or is it simply not a priority? Does any AI researcher or practitioner actively use the ethics guidelines in their work? Does any AI researcher or practitioner actively use the technical solutions that exist? It is necessary to know where the problem lies in order to find which lever to pull to correct for it.

The literature is rife with examples and case studies of gender discriminating AI algorithms, as mentioned in chapter 2.2.3. However, when looking for literature, little to no topology or model was found for the issue of gender bias. How does gender bias enter the AI system? An overview of the issue might make it easier to address. A model of how gender bias might enter the AI system might make the assessment for gender equality in an AI system more systematic and easier to approach. Additionally, literature comparing the importance and impact of causes and solutions has not been find.

3 Theory

Feenberg's philosophies of technology (2003, 2006) have been chosen as a framework to understand and analyze the perspectives and understandings of technology of the informants. It is relevant to understand and dissect their technological understandings because they might be shaping the informants' actions and perspectives on gender bias in AI. The philosophies of technology will make it easier to connect observed traits with the informants' beliefs and practices. This thesis focuses on the *instrumentalist* perspectives and *Critical Theory of Technology*, which will be explained in the next sections.

As will be explained in Chapter 4 – Research Approach, this study uses a critical research approach to investigate gender bias in AI. Feenberg's *instrumentalization theory* is especially useful to critically analyze how aspects such as society, culture, power, assumptions, beliefs, and practices influence the development of AI systems. An understanding of how AI systems are shaped by the aforementioned aspects may contribute to an explanation as to why AI systems sometimes become gender biased.

Several research articles have called upon the utilization of gender theory to understand and address gender bias in AI (Draude, Klumbyte, Lücking, & Treusch, 2019; Leavy, 2018). Therefore, feminist theories and concepts such as *gender equity*, *Standpoint theory*, *Design Justice Principles*, and *intersectionality* are used to understand the informants' perspectives on gender bias in AI. The gender theory also provide a better comprehension of why gender inequalities occur in AI systems and how to address them.

Causes and solutions are commonly listed and explained in research literature without a comparison or indication of which causes or solutions would be of greatest impact or importance. Furthermore, the literature rarely provides any differentiation between which lever of intervention will be harder or easier to implement. Meadows' (1999) Leverage Points Theory is for that reason used to categorize the reported causes and solutions for impact. Ease of implementation is also indicated because, as the section on Leverage Points Theory will explain, the degree of difficulty increases with impact.

3.1 Different Perspectives on Technology and Designers

3.1.1 Philosophies of Technology

Below is a summary of Andrew Feenberg's Philosophy of Technology (2003, 2006). Feenberg proposes the following diagram of different views of technology:

Table 1 *Feenberg's Definitions of Philosophies of Technology*

Technology is:	Autonomous	Humanly Controlled
Neutral **(complete separation of means and ends)**	**Determinism** (technology control society, optimistic)	**Instrumentalism** ("Guns don't kill people, people kill people.")
Value-laden **(means form a way of life that includes ends)**	**Substantivism** (means and ends linked in systems, critical, dystopia)	**Critical Theory** (choice of alternative means-ends systems)

Note. The two axis of the diagram represent how different views see the autonomy and neutrality of technology. Adapted from *"Defining Technological Literacy Towards an Epistemological Framework",* by Feenberg, A., 2006, p. 10, New York, NY: Palgrave Macmillan US.

Although the diagram divides different views into neat boxes, the views are not clearly demarcated, and a person's view might overlap several boxes. Further explanation of the table's axes and boxes follow below.

Vertical axis: Neutral vs. Value-laden

The vertical axis of the diagram represents the differentiation of whether technology is seen as neutral or value-laden. Is technology simply a collection of mechanisms strung together without any intentions or values of its own? Feenberg's view is that technology can contain value the same way a banknote can hold value. It is not any specific physical property of a banknote that makes it valuable, but that it has its own way of containing value as a social entity. We, as a society, ascribe the value to the banknote. Although there is not necessarily any code line that could be pointed to as value-laden, the sum of its parts still holds a value. An example of a such value could be efficiency or power.

Horizontal axis: Autonomous Technology vs. Humanly Controlled

The horizontal axis represents the level of control humans have on the direction technology develop. Do humans decide where the technology will go? Are we able to align the next step of development with our intentions? Or has the evolution of technology gotten an autonomous life of its own? Instrumentalism, in addition to believing that technology is neutral, believes that humans control how it will develop.

Instrumentalism

According to Andrew Feenberg the dominating view of modern society is that means and ends are separated. Looking at the saying: "Guns don't kill people, people kill people", he says that most people think of means and ends as separate entities. A gun is neutral, but can be used for good when the police is using it to fight crime, but it can also be used for bad when someone is robbing a bank. This type of liberal faith in progress is called instrumentalism.

The instrumentalist views technology as a neutral tool or instrument that humans can use for their own needs. Their view is that this neutral tool can be used to execute a pre-existing value more efficiently. For instance, an emergency response operator who uses a caller's GPS location to send an ambulance because they value human life.

Determinism

Feenberg says that determinists agree with instrumentalism on technology being neutral (2003). However, technological determinism presents the view that humans don't control technology, but that technology controls society. Technology influences society to focus on progress and to be more efficient. Humans don't adapt technology to align with our intentions, on the contrary, humans have to adapt to technology as it progresses, he says. When technology becomes faster, humans must adapt to the new speed (Quan-Haase, 2013). Despite the lack of control, determinism according to Feenberg is presented as optimistic (2006). An example of this view in AI is that AI is here to stay and that it will increasingly shape all areas of our lives in the future, but humans must simply find a way to comply with this progress.

Substantivism

Substantivism does not view technology as neutral and attributes substantive values to technology. A substantive value includes the commitment to a specific idea of what constitutes a good life. Substantivism believes that the values in technology cannot be chosen by humans, but are ingrained in the technology. Using technology is a value choice and it can't be used for a purpose that contradicts the value of the technology. For instance, a gun is made to kill so the use of a gun is colored by that value regardless of the intentions of the user.

Substantivism compares the use of a technology to subscribing to a religion. By using technology to increase the efficiency of one's life, a different way of life has been chosen. This choice entails the exclusion of contradicting values, the same way converting to a religion would. For instance, using a gun would mean the rejection of the notion that all lives are of equal value, regardless of the purpose. A technological society would automatically be shaped by the values of technology such as efficiency and power. Substantivism is critical about the future of technology. In the most extreme cases, it imagines a dystopic future where humans have become the submissive cogs in the machinery of technology. An example of this view in AI, is the idea that AI will one day become so sentient and autonomous that humans will lose control and be eradicated by the AI.

Critical Theory of Technology

Critical Theory of Technology (CT) sees technology as value-laden and humanly controllable, traits that are shared with substantivism and instrumentalism, respectively. It recognizes the values in technology and its potential harmful consequences, but it believes that those values can be changed by humans. Although human control is possible technology is still not seen as a neutral tool. It sees technology as having an additional nuance to the efficiency that technology brings to life. This nuance is the difference between efficient weapons and efficient medicine. Technology is seen not as tools but as frameworks for our lives. However, unlike substantivism it sees technology as a way to frame *more* than one way of life. We can then use legislation to choose whether we want our way of life to be a world with or without guns.

Feenberg refers to this process as *democratic interventions* of technology. A democratic intervention does not necessarily entail the public voting over different technological devices, but rather protesting or innovating technology to promote change. Feenberg suggests protesting unwanted changes, like a nuclear power plant, or innovating an existing technology, like how advanced internet users created email. He believes it is our failure to create sufficient institutional human controls over technology that leads to problems.

Feenberg says that there is a trend towards greater participation in design. He further argues that this participation and extension of democracy is the only way to save us from “certain destruction”. This trend needs to continue until citizenship means a right to exercise control over the technological framework of our lives, the same way citizenship means the right to exercise control over the laws that govern our society.

A CT perspective presents the view that tradition and culture influence design practices. The *Instrumentalization Theory* developed by Feenberg can then be used to investigate the cultural values and practices that surround the technology of AI (Feng & Feenberg, 2008). In this perspective, power is seen as located on the macro-level of traditions and culture and not solely on the micro-level with the actors.

3.1.2 Instrumentalization Theory

In the book section *Thinking about Design: Critical Theory of Technology and the Design Process*, Feenberg proposes something called *Instrumentalization Theory* (InT) (Feng & Feenberg, 2008). InT is not to be confused with Instrumentalism outlined in the chapter above. InT is an approach for analyzing what *cultural resources* impact a design. The Instrumentalization theory has the perspective that the design process is not only shaped by the interest and will of the actors but is also molded by the background, tradition, and culture that exists within the society it is designed. *Underdetermined design choices* are design choices that are unconsciously made based on this background. (Feng & Feenberg, 2008)

The background in InT consists of two levels:

- Level 1: Basic technical operations, taken for granted beliefs, and everyday practices.
- Level 2: Current power relations and socio-cultural conditions. A “history of technical choices” that affects the “culturally biased knowledge” (Feng & Feenberg, 2008, p. 112)

Feenberg divides Instrumentalizing objects into a device into two processes:

- Primary Instrumentalization: Simplifying objects to understand their function so that they can become the technical elements that make up a device.

- Secondary Instrumentalization: Designing neutral technical elements into a strongly biased device to fit a particular social context. The "social design of society" has a big impact on the most important parts of a device (Feng & Feenberg, 2008, p. 114).

Feenberg defines technical heritage as the assumptions, practices, and world perspectives that are inherited from years of research building on top of each other within a field. This technical heritage decides what is considered technically feasible and has a big influence on designs because they lead to underdetermined design choices; already dismissed options and taken for granted assumptions. The technical code are standards that come from social requirements and standard ways of perceiving a device of technology (Feng & Feenberg, 2008).

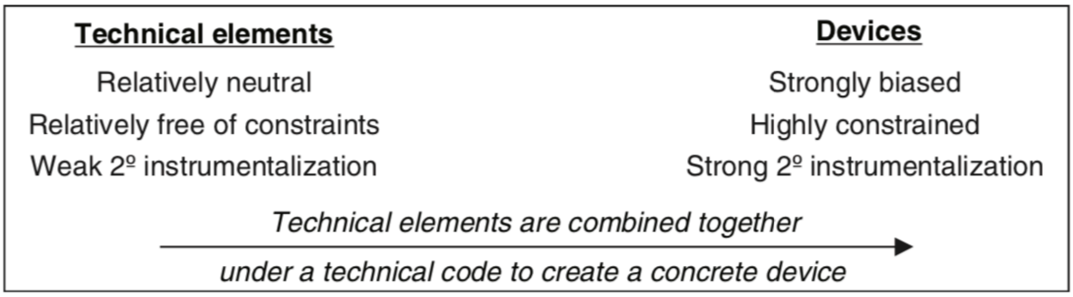

Technical elements
Relatively neutral
Relatively free of constraints
Weak 2º instrumentalization
Devices
Strongly biased
Highly constrained
Strong 2º instrumentalization
Technical elements are combined together
under a technical code to create a concrete device

Figure 3. Relationship between technical elements and concrete devices reprinted from Feng and Feenberg, 2008, p. 114.

3.1.3 The Power of the Designer According to Instrumentalization Theory Perspective

In the book chapter *Thinking about Design: Critical Theory of Technology and the Design Process* Feng & Feenberg writes about different ways of viewing the designer (2008). Below is a summary of those views. Although the referenced book chapter is about the power of the designer, the practitioners interviewed for this master's thesis project can be seen as designers since AI systems are often designed by engineers, researchers, and developers, such as them.

If Designers Are Seen as Powerful

Norman (1988) said that bad design is due to engineers, programmers, and managers doing the design work instead of designers (p. 156). He meant that the solution for better design was to create more enlightened designers. Norman also said that "The designer must assume that all possible errors will occur" Norman (1988, p. 36).

The view of a powerful designer is related to instrumentalism where the role of the designer is to develop technology to requirements that are external to the design process, and the design process is seen as a mainly technical process.

If Designers are Weak or Inconsequential

Factors that affect the power of the designer according to Feng & Feenberg (2008): Economic, political, institutional, social, and cultural factors. Each of these aspects diminish the autonomy of the designer. There is for instance a conflict between the corporate interests, norms and culture, and the intentions of the designer and AI practitioner (Feng & Feenberg, 2008).

Feng & Feenberg (2008) further writes that even the "purely technical" activities are never unaffected by value-laden and cultural rules. "As numerous STS studies have shown, the making of such standards are as much political as they are technical in nature: technical standards are never "purely technical" "(Bowker and Star, 2000).

Feenberg (Feng & Feenberg, 2008) states that design always exhibits social bias because it is inherent in optimizing. By choosing what to optimize for the designer chooses what to prioritize and this prioritization is the social bias. An example is to whether optimize for the cost of the manufacturer or the environmental cost on the planet.

3.2 Gender Theory and Feminism in AI

Jule (2014, p. 2464) defines gender theory as "the study of what is understood as masculine and/or feminine and/or queer behavior in any given context, community, society, or field of study." Feminism is defined as a political social movement that brings gender oppression into the open, connects it to politics, structures, institutions, and rally people to remove gender-based oppression by transforming gender relations (Motapanyane, 2014).

Susan Leavy (2018) and other researchers (Draude, Klumbyte, & Treusch, 2018) argue that it is necessary to include gender theory in order to understand how AIs can become biased. One of the examples Leavy puts forth is how language translation models are affected by the gendering of words in different languages, this is also mentioned in Chapter 2.

Theories and concepts from gender theory and feminism have therefore been chosen to be included as theories for this thesis. The chosen theories and concepts will be outlined below: Equity, Sandra Harding's standpoint theory and *strong objectivity*, *Design Justice* principles, and *intersectionality*.

3.2.1 Gender Equality and Gender Equity

Gender equality and *gender equity* are two terms that are sometimes mixed up but they are quite different. **Gender equality** refers to how men, women, and other gender identities should be treated equally and have equal opportunities. It is also called *formal equality* (World Health Organization, 2011). **Gender equity**, on the other hand, is defined as considering the specific circumstances and needs of men, women, and other gender identities, and treat them differently in order to create equal opportunities and equality of results (World Health Organization, 2011). Gender equity is also sometimes called *substantive equality* (World Health Organization, 2011).

Many places where AI and ethics is discussed, the term "fair AI" is mentioned. It is unclear what characteristics make up a "fair AI" and whether that should be the goal. Two researchers argue that rather than a fair AI, we should be creating AIs with not just equality, but equity. The issue with fairness and equality is that it does not acknowledge or balance for the historical injustices that has accumulated over time (D'Ignazio & Klein, 2020).

For instance, a resume filtering AI that treats all resumes the same might penalize women for having gaps in their resume resulting from maternal leave if it interprets gaps in the resume with a less ideals hire. A more equitable AI could adjust a scoring system in a way that does not penalize women for going on maternal leave. Gender equity is a concept from gender theory, and as Leavy and Draude et al. say, gender theory is needed in machine learning and is therefore used in this thesis (2018).

3.2.2 Standpoint Theory, Strong Objectivity, & Design Justice

Standpoint theory originates from the 1970-1980s feminism and feminist Sandra Harding coined the emergent concept of strong objectivity (1995). Standpoint theory claims that what we do in social relations enables and limits what we can know (Harding, 1995). Harding says that all knowledge is partial and is limited by its historical point in time and taken for granted cultural assumptions (1995). She continues that these assumptions can only be pointed out by someone who is outside of the culture (Harding, 1995).

The concept of strong objectivity comes from feminist standpoint theory (Harding, 1995). Harding sees neutrality as an obstacle to maximize objectivity because interests and values can distort the knowledge (1995). She explains that the ideal of neutrality is harmful because powerful institutions and groups can hide behind this objectivism and defend their political practices by claiming to be value-neutral when they are not (Harding, 1995).

Standpoint theory claims that "knowledge is always socially situated" (Harding, 2004, p. 7). According to Harding, only from the marginalized standpoint can one see both the views of the privileged *and* the disadvantaged. Harding believes that the only way to strengthen objectivity is to create knowledge from the standpoint of the marginalized. Research should therefore start from the point of those who are excluded and marginalized by the current framework (Harding, 1995). Because of the limited viewpoints of individuals it is the communities and the collection of voices, not individuals, who produce knowledge (Harding, 1992). A collection of voices that includes those who are disadvantaged strengthens objectivity, according to Harding.

Design Justice Network Principles

The Design Justice Network Principles are a set of 10 principles that are based on similar ideas as Standpoint Theory (Design Justice Network, 2018). The network states that the principles aim to center the voices of those who are the most marginalized in a collaborative design process. They were created by 30 designers and individuals at the Allied Media Conference in 2015 (Costanza-Chock, 2018). The most relevant principles for this thesis are highlighted in bold below.

1. **We use design to sustain, heal, and empower our communities, as well as to seek liberation from exploitative and oppressive systems.**

2. **We center the voices of those who are directly impacted by the outcomes of the design process.**

3. **We prioritize design's impact on the community over the intentions of the designer.**

4. We view change as emergent from an accountable, accessible, and collaborative process, rather than as a point at the end of a process.

5. We see the role of the designer as a facilitator rather than an expert.

6. **We believe that everyone is an expert based on their own lived experience, and that we all have unique and brilliant contributions to bring to a design process.**

7. We share design knowledge and tools with our communities.

8. **We work towards sustainable, community-led and -controlled outcomes.**

9. We work towards non-exploitative solutions that reconnect us to the earth and to each other.

10. Before seeking new design solutions, we look for what is already working at the community level. We honor and uplift traditional, indigenous, and local knowledge and practices.

(Design Justice Network, 2018)

Intersectionality

An intersectional approach to solve gender bias in AI is important to not implicitly privilege white women (Costanza-Chock, 2018; S. M. West et al., 2019). Algorithmic bias audits are flawed because they only look to correct for one variable (Costanza-Chock, 2018). An intersectional view of bias audits would mean to check not just for bias against women, but also for bias against women with other vectors of discrimination. For instance colored women, women from underprivileged backgrounds, women from a lower socio-economic status, women with disabilities, etc. Without an intersectional view, solutions will not be able to address the burdens of those who are multiply-burdened (Costanza-Chock, 2018).

3.3 Leverage Points Theory

In the journal article *Leverage Points – Places to Intervene in a System*, Donella Meadows (1999) outlines a taxonomy that categorizes different types of interventions according to their level of impact. This theory is situated within the Systems Thinking perspective and Meadows made it with the intention of helping others recognize levers of intervention in a system. The theory of Systems Thinking provides a perspective of considering the entirety of a system when trying to change it (Kim, 1999).

The Leverage Points (LP) theory by Meadows (1999) are summarized below in the order from the most impactful interventions to the least impactful. In her article, she uses a bathtub as a metaphor for a system. The state of the system is the level of water and the inflow and outflow of water are the factors that can change the state of the system.

Meadows specifies that the order of the list is not absolute and that all LP have exceptions that might belong further up or down the list. She further warns that the higher a leverage point is, the more the system will resist the change.

1. The power to transcend paradigms

Meadows (1999) points out that the highest lever is the understanding that no paradigm is an eternal panacea. She says that all beliefs and paradigms are limited and cannot fully explain the complexity of the world, and it is therefore better to remain open-minded and to be able to adopt alternative paradigms instead of clinging to the existing ones.

2. The mindset or paradigm out of which the system—its goals, structure, rules, delays, parameters—arises

Meadows describes the paradigm of societies as the entrenched beliefs that are so obvious and innate that they are not spoken. Examples of such beliefs are "growth is good" or that it is possible to "own land" (1999, p. 17). The beliefs and paradigms of society are the sources of which the different parts and levers of the system spring out from.

The paradigm is therefore a very high leverage point, however, Meadows warn that it is the most difficult one to change. On the other hand, they do not have physical limitations, they are not expensive to change, and they can be quick. A paradigm shift can be swift in an individual but they are stubbornly resistant for societies.

Meadows says that one must continually point out the system's flaws and irregularities to spark a paradigm shift, and to model the system so that one can step outside the system to see it. She advises that one should ignore those who are strongly oppositional and instead focus on those who are open to change. Another intervention is to put people who understand the new paradigm in places of power and visibility.

3. The goals of the system

As mentioned in leverage point 12, changing the parameter of who is hired is a low-level intervention. The exception is if a new person at the top has the power to change the goals of the system. Meadows (1999) describes the goals of the system as a very high leverage point because all lower leverage points will be changed to meet the goal. However, she differentiates between system goals and individual goals. The goals are not what anyone *says* but what the system actually *does*: Actions such as survival, evolution, resilience, or expansion.

Meadows argues that the system goal of corporations are not profit because profit is just a necessity to remain in the game of business. She says the real question is: What is the *point* of the game? She argues that the system goal of corporations is to "engulf everything" and that it is imperative to have negative feedback loops that are stronger if the market is to remain competitive (1999, p. 17).

4. The power to add, change, evolve, or self-organize system structure

This LP refers to the ability to create new structures and behaviors. For instance, evolution in nature or technical advances and social revolutions in society. This LP represents the *ability to change* any aspect of the system from LP 5 to 12, and the power to write the rules of how the system can self-organize.

According to Meadows, the accumulated knowledge is the raw material of technological evolution, whereas variety in inventions comes from creativity. The market and funding structures are the selection mechanisms that decide what designs get to survive. If a system prevents experimentation or learning, and instead stops the evolution on a homogenous culture, it will be less resilient and eventually collapse.

5. The rules of the system (such as incentives, punishments, constraints)

The rules define the limits of the system and the extent of freedom within the system. Meadows says that "power over rules is real power" because it is a high leverage point that greatly affects behavior (1999, p. 14). She writes that looking at the rules and who has power over them can reveal the most innate malfunctions of a system.

For examples, Meadows describes a meeting where a global trade regime was made by and for corporations where the rules would lead to positive loops of "success to the successful". Its lack of outside feedback, transparency, and accountability would have led to a race to the bottom between nations, because they would compete for companies' business by weakening rules meant to protect the environment and society.

6. The structure of information flows (who does and does not have access to what kinds of information)

Create a new loop where information goes to new places and causes behavioral change. An example of this leverage point is where companies who are exposed of nefarious practices change behavior because of the risk of losing customers. Sharing of previously missing information can restore or create a feedback loop that can prevent a system failure. Meadows writes that this is a powerful LP that can lead to accountability of those in power.

7. The gain around driving positive feedback loops

Positive feedback loops are systems where the more you have of something, the more likely you are to get more of it. For instance, the more political or economic power an organization has, the easier it is for that organization to gain more of that power. They are "success to the successful" loops (Meadows, 1999, p. 12). According to Meadows, positive feedback loops are the causes of growth but also the cause of collapse of systems. She states that a system will eventually ruin itself if it has an uncontrolled positive feedback loop.

Systems where positive feedback loops are much faster than negative feedback loops can turn into chaos. The system can recover if the growth rate is slowed so that delayed negative feedback loops have time to catch up. An example of this is the positive feedback loop of COVID-19 that hurled the world into chaos. By slowing the infection rate through measures such as lockdown, lives could be spared while the medical companies worked to make a vaccine.

8. The strength of negative feedback loops, relative to the impacts they are trying to correct against

Meadows argues that a negative feedback loop requires: 1. A goal, 2. a way to monitor the state, and 3. a response mechanism. The strength of the negative feedback loop is determined by how fast and accurate the monitoring system is and how fast and powerful the countering mechanism is.

Markets are controlled by supply and demand, and a change in price is the negative feedback that can change the demand. Democracy is another example of a negative feedback loop: Voters can decide which politicians have the power depending on the actions of the politicians. However, this negative feedback loop relies on an informed population, and fake news or unclear information can weaken this feedback loop. This negative feedback loop is further weakened by what Meadows calls "the brainwashing power of centralized mass communications". This makes the correction weak relative to the negative impact of the politicians that are elected through confusion. Leveling the playing field is necessary in order to weaken the power of those trying to weaken the negative feedback loop.

Examples of interventions that strengthen negative feedback loops are the Freedom of Information Act that increases government transparency, systems for monitoring and reporting on environmental damage, protection of whistleblowers, and pollution taxes.

9. The lengths of delays, relative to the rate of system change

This LP refers to delays in feedback loops, such as the time it takes from a technology is installed until the effects on society is seen or how long it takes for a price to adjust to an imbalance in market-demand. Longer delays of feedback can cause interventions to overshoot or undershoot. This is similar to how a shower can be too hot or too cold if it takes a lot of time for the water to change its temperature after the faucet has been adjusted.

Too short delays can cause overreaction, too long delays can cause collapse if there is a point of no return from irreversible damage. An example of irreversible damage is the extinction of certain species. However, it is difficult to change the length of delays. For instance, it is probably impossible to change how long it takes for pandas to grow and start mating. Meadows says that it is therefore easier to slow the change rate of the system so that there is enough time for the feedback to be seen. She says that it is easier to slow the economic growth (leverage point 7) than to speed up the development of technology.

10. The structure of material stocks and flows (such as transport networks)

Physical infrastructure like roads in a country or the plumbing of the bathtub. This leverage points is low because it is slow and expensive to change physical infrastructure. In order to use this as a leverage point, it is necessary to design the system well from the beginning.

11. The sizes of buffers and other stabilizing stocks, relative to their flows

A system can be stabilized with a bigger buffer (bigger bathtub) or become slow to change if the buffer is too big (swimming pool with bathtub faucets). Examples are the inventory of a store or savings in the bank.

12. Constants, parameters, numbers (such as subsidies, taxes, standards)

The equivalent of "diddling with the details, arranging the deck chairs on the Titanic" (Meadows, 1999, p. 6). This LP is similar to adjusting the faucet of the bathtub. Companies can change their prices and employee salaries to adjust the level of their profit bathtub. However, Meadows categorizes this as a low-level intervention because parameters do not tend to change behavior. Hiring different people to adjust the same faucets is likely to create similar results. She adds, the exception is if the parameter has the ability to pull a more impactful lever.

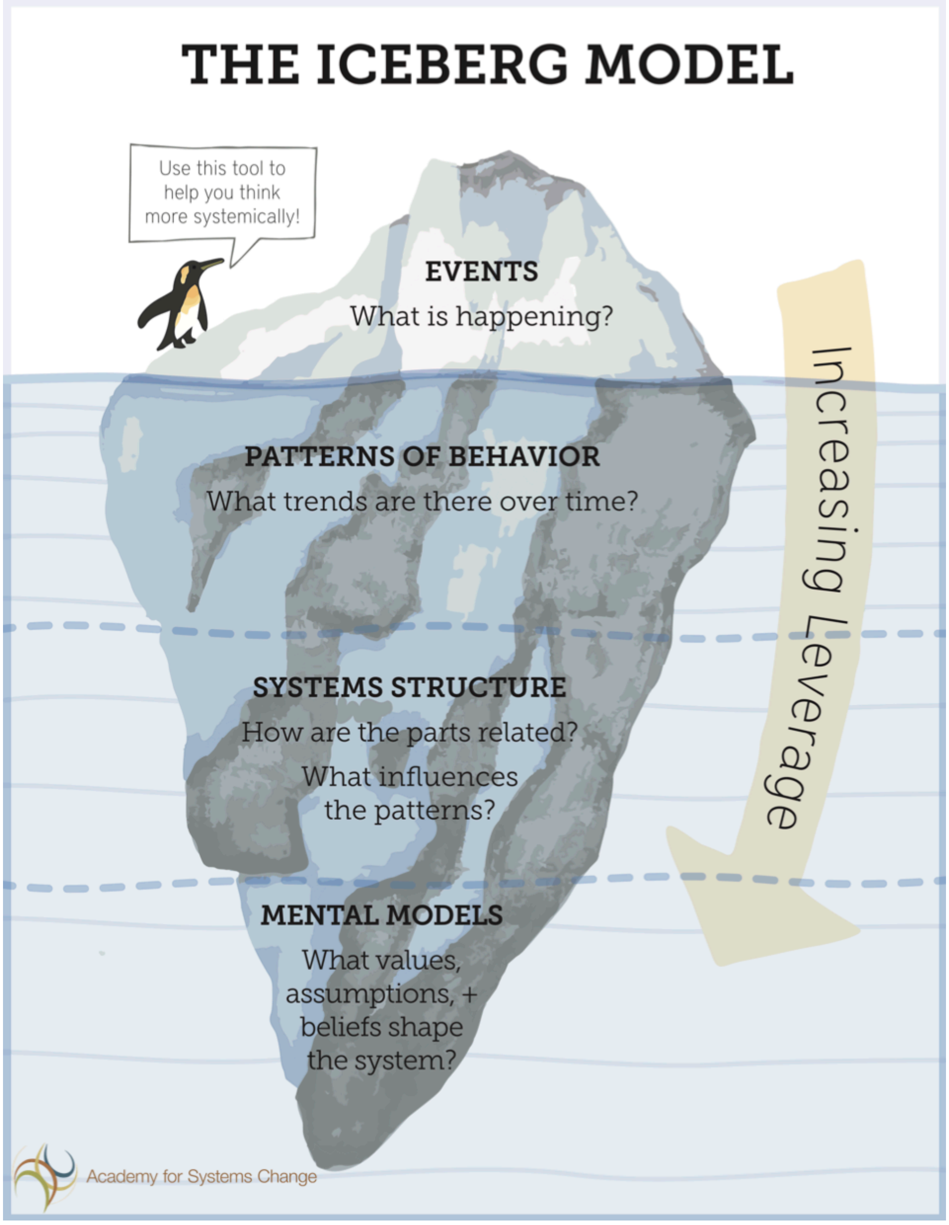

THE ICEBERG MODEL
Use this tool to help you think more systemically!
EVENTS
What is happening?
PATTERNS OF BEHAVIOR
What trends are there over time?
SYSTEMS STRUCTURE
How are the parts related?
What influences the patterns?
MENTAL MODELS
What values, assumptions, + beliefs shape the system?
Increasing Leverage
Academy for Systems Change

Figure 4. Meadows' Iceberg Model of the possible causes of certain events. This model is similar to Leverage Points Theory, but is more condensed. The model is used to aid thinking more systematically. The deeper a level is, the higher leverage and impact does the cause have. Reprinted from Systems Thinking Resources, in *Academy for Systems Change*, n.d., retrieved

April 23, 2021, from http://donellameadows.org/systems-thinking-resources. Licensed under Creative Commons, reprinted with permission.

Table 2 (below) organizes the current state outlined by the literature review and future solutions suggested by literature into the LP theory of Meadows. Examples that are listed by Meadows that *could* be relevant for the AI field are also listed. This table is useful for understanding the LP theory in the AI context and for comparing the impact of different solutions.

Table 2
Table of Current State and Future Solutions from Literature Sorted into Leverage Points

Leverage Points	**The current state according to literature**	**Solutions for the future from literature**
12. Constants, parameters, numbers (such as subsidies, taxes, standards)	Lack of diversity.	More diverse hires.
11. The sizes of buffers and other stabilizing stocks, relative to their flows.		
10. The structure of material stocks and flows	Oppressive AIs are already in use everywhere around us and affecting us every day.	
9. The lengths of delays, relative to the rate of system change		
8. The strength of negative feedback loops, relative to the impacts they are trying to correct against	The AI Industry lobbying and skewing research in their favor the same way the tobacco industry did.	Algorithmic audits. E. g. Protection of whistle-blowers, global government, truth in ads (Meadows, 1999).

7. The gain around driving positive feedback loops	The rapid growth of Big Tech erodes democracy and exacerbates injustice and inequities. Those in power gets more powerful. The Global AI Race against other big nations.	
6. The structure of information flows	Algorithms are opaque due to being protected as intellectual property.	Transparency Accountability
5. The rules of the system		Laws regulating AI. Ethics guidelines.
4. The power to add, change, evolve, or self- organize system structure Changing any aspect of the system from Leverage Point 5 to 12.	The AI Revolution is pulling the lever in the wrong direction for gender and racial equity.	
3. The goals of the system	Capitalism -> Survival -> Profit -> Power -> Monopoly (Meadows, 1999).	
2. The mindset or paradigm out of which the system—its goals, structure, rules, delays, parameters—arises		Intersectional perspective. Books, documentaries, awareness, activism. Education
1. The power to transcend paradigms		

4 Research Approach

Myers (1997) explains, the critical paradigm assumes that the current social reality depends on the previous history that has been created by other humans. This fits well with the critical theory of technology used for analysis.

Myers (1997) further describes the critical paradigm as a philosophical perspective that recognizes that although people might have the freedom to improve their environment, they are still limited by the social, cultural, and political constraints of their circumstances. This is similar to Feng & Feenberg's Critical Theory of Technology (2008), which recognizes that designers are constrained by the power structures of their circumstances, and using CTOT as a theory is in line with Myers & Klein's first principle (below). Myers & Klein (2011, p. 25) propose the following principles for critical research:

1. The principles of using core concepts from critical social theorists
2. The principle of taking a value position
3. The principle of revealing and challenging prevailing beliefs and social practices
4. The principle of individual emancipation
5. The principle of improvements in society
6. The principle of improvements in social theories

A critical paradigm for the research project has been chosen due to the aim of social critique and improvement of the issue investigated. The aim of social critique and social improvement fits well with principle 2, taking a value position; and principle 5, improving society. The value advocated in this thesis is equity. The intent behind exploring RQ3, where gender bias comes into an AI system, is to contribute to the discovery of potential solutions. The thesis also does not have a hypothesis to test, which would have been more in line with a positivist paradigm (Myers, 1997).

The critical perspective focuses on the contradictions and conflicts of society, and it attempts to create release from the dominant power structures (Myers, 1997). In this case, the focus is on the conflict of the pressing issue of gender bias in AI and how AI practitioners view and act towards that issue. Following principle 4, the aim of this thesis project is to contribute towards the emancipation of people, especially women, from the dominating gender bias in AI systems. The main RQ, RQ 1, and

2 attempts to follow principle three by exploring perspectives on gender bias in AI (GBAI), the understandings of technology and the ethics practices of AI practitioners.

4.2 Methodology & Methods

Exploratory Qualitative Research

The aim of this study is to map and describe the perspectives on gender bias in AI amongst AI practitioners in Norway. The study limits its geographical scope to only explore the views of practitioners in Norway. An exploratory qualitative research methodology was chosen, due to not knowing in advance what will be found, to keep an open approach, and because there is no clear hypothesis to test. Qualitative research is research that leads to descriptive data, it has a focus on people's thoughts and behaviors, and the researcher approaches the subject with a holistic view of the people or environments studied in a way that does not reduce information to quantified numbers (Taylor, Bogdan, & DeVault, 2016).

The goal of the study is to deep dive into the case studies of the participants to uncover insights that otherwise would have been unattainable using a survey. In contrast to quantified data, the upside of a qualitative approach is that qualitative data contain the social and institutional context that is attached to the views of an informant (Myers, 1997). Since this thesis wants to investigate where gender bias might enter an AI system and power issues might be relevant for that investigation, the social and institutional context might be especially relevant.

Yin (2009) defines three conditions for choosing a method: type of research question, whether the researcher has control over the behavioral events, and to which extent the focus is on historical or contemporary events. According to Yin (2009), some research questions of the type *what* are exploratory and for exploratory research all research methods can be useful. *Multiple case studies* have been chosen as a method because they are particularly suitable for exploring current phenomena in detail within a real life context, especially when it is unclear where the phenomenon ends and the context begins (Yin, 2009). Case study research often uses multiple sources of data collection like documents and interviews to collect empirical evidence on a certain topic (Myers, 1997). This study conducts several semi-structured interviews.

Case studies are not statistically generalizable because those studies does not represent a sample. Instead, the goal is to broaden and generalize theories to achieve analytic generalization (Yin, 2009). Theories are therefore used to guide data collection and analysis by providing a hypothetical explanation as to why actions, events, and thoughts occur (Yin, 2009). The cases in multiple case study research are seen as independent cases, and patterns in subsequent replicated cases can support or disconfirm the findings of the previous cases (Robson, 2002).

4.2.1 Semi-structured interviews

Semi-structured interviews are the most common type of interviews in qualitative research in information systems (Myers & Newman, 2007). Semi-structured interviews have been chosen to allow the AI experts in Norway lead the way to insights, rather than pre-defining a script that eliminates improvised exploration. Semi-structured interviews are more suitable for exploratory research (Robson, 2002). A survey would be difficult to conduct because it is not known what the right questions to ask would be and surveys do not facilitate follow-up questions for additional probing. A literature review alone would be somewhat limited because this is a new area of study. A content analysis of the websites or ethics documents of an organization would require that organizations had documented ethics practices, if they had any. Interviews have therefore been chosen for this study.

Implementation: Interview Guide

A semi-structured interview guide was made in advance. Due to the challenge of getting participants to be interviewed the interviews were between 30 to 60 minutes long in order to increase the likelihood of more subjects participating (Robson). The chosen interview questions were informed by the literature review which shaped which questions were included. Notes were written during the interview to make sure that relevant follow-up questions were not forgotten during the interview, and to have an easily available "summary" of the interview. The topic of gender bias in AI was generalized to *ethics in AI* to decrease the likelihood that participants would do research on the topic in advance and answer in ways to make them look good or please the interviewer (Robson, 2002).

Implementation: Researcher Behavior

Things that were done to increase participants' level of trust and comfort: Nod, agree, smile. Emphasize the anonymity and their rights. Start questioning broadly and try to let them broach the topic of gender bias on their own. The participants were offered the opportunity to read and provide feedback on a draft of the thesis to ease their concerns about what was said and to increase their level of trust (Taylor et al., 2016).

4.2.2 Informants

The sample of informants was a convenience sample due to the constraints of resources and lack of contacts in the field. The participants were contacted through the network at the University of Oslo, Oslo Metropolitan University, through personal contacts, and through referrals from interviewed participants using snowballing. Some participants were also contacted through professional networks like Nova (Nova, 2021) and ODA (ODA, n.d.). Others were contacted because of their speaking engagements on relevant AI events.

In order to limit the focus of the thesis and to avoid results being clouded by cultural differences, the subjects for this thesis has been chosen to be AI practitioners who mainly work or reside in Norway. Subjects were also chosen based on how close they were to the actual AI development, when that was possible. For instance, an engineer would be preferred over a department leader. This choice was made based on the assumption that engineers know more about the practical elements of creating an AI, such as how data is processed prior to training. Since this thesis aims to explore the ethics practices related to the programming, development, and life cycle of AI and gender bias, it made sense to focus on the engineer.

Implementation: Informant Selection

The group of informants is a convenience sample with the aid snowballing. This thesis has defined the different participant sectors as is shown in Table 3 below.

Table 3
Definitions of Sectors

Sector name	Abbreviation used as prefix code on participants	Definition
Academic Research	AR	Researchers with a focus on AI in academic institutions like universities.
Government Institution	GI	Employees at institutions that belong to the government or whose organization receive significant funding from the government.
Private Company	PC	Employees and data scientists who belong to a mid-size to big private company.
Startup	SU	Developers and founders of a company without a finished product or a product that is not yet mass-

		adopted. Has maximum 20 employees.

This project aims to interview equal numbers of men and women in case someone's gender might affect the level of awareness about gender bias. Due to the potential bias due to one's gender regarding the topic, a gender balance was strived for in order to account for a potentially skewed result if only one gender was interviewed. However, the acquired participants are limited by the researcher's network and access to potential informants. A complete gender balance was not achieved with the 5 female, 7 male, and 1 'other' participants in the study. However, a ratio of 46% participants who are not cis-male is far more balanced than the current gender ratio in IT which is estimated to be 20% female in Norway (Brombach, 2016).

4.2.3 Thematic Analysis

The interviews were analyzed using *thematic analysis* (TA) following the guidelines of Braun & Clarke (2006) and Braun, Clarke, Hayfield, & Terry (2018). The field of TA, as Braun et al. describes it, has a history of being used to analyze qualitative data which is suitable for this qualitative study. Within TA there are different types and in this thesis the approach *Reflexive TA* has been chosen for its view on researcher subjectivity as a resource.

According to Braun et al., the goal is not to depict the data as "correct" as possible, but to form a compelling story based on the data for the chosen agenda of social change. The objective of Reflexive TA is to interpret the data through the lens of the researcher's position in society, their theoretical beliefs and ideologies, rather than to try to "objectively" summarize data because they do not think that such objectivity is possible. (Braun et al., 2018)

Furthermore, a strength of Reflexive TA is its iterative process, as opposed to having a strict codebook from the start (Braun et al., 2018). Although some domain summary type of themes were coded, a deeper analysis is necessary to map the unspoken patterns. Braun et. al define *domain summaries* as summaries of what participants have said regarding a specific topic without distinguishing whether the replies contradict each other (2018).

Considering that the main research question is "What perspectives on the issue of gender bias in AI are there among AI practitioners in Norway?", it is very fitting that Reflexive TA suits purposes such as to portray lived experiences and investigate causes and views related to a phenomenon (Braun et al., 2018). The six phases of thematic analysis are described in Table 4.

Table 4
Phases of Thematic Analysis (Braun & Clarke, 2006, p. 87)

Phase	Description of the process
1. Familiarizing yourself with your data:	Transcribing data (if necessary), reading and re-reading the data, noting down initial ideas
2. Generating initial codes:	Coding interesting features of the data in a systematic fashion across the entire data set, collating data relevant to each code
3. Searching for themes:	Collating codes into potential themes, gathering all data relevant to each potential theme
4. Reviewing themes:	Checking if the themes work in relation to the coded extracts (Level 1) and the entire data set (Level 2), generating a thematic 'map' of the analysis
5. Defining and naming themes:	Ongoing analysis to refine the specifics of each theme, and the overall story the analysis tells, generating clear definitions and names for each theme.
6. Producing the report	The final opportunity for analysis. Selection of vivid, compelling extract examples, final analysis of selected extracts, relating back of the analysis to the research question and literature, producing a scholarly report of the analysis.

4.2.4 Test Interviews

The two test-interviews in Table 5 were conducted in June and September 2020. The interview guide was edited after the test interviews. Questions were re-phrased and some were removed to increase the likelihood that there would be time to ask all or most questions during the timeframe of 30 to 60 minutes. For instance, the question "How does your organization approach AI and ethics?" was removed to decrease the number of questions about their organization to make room for questions regarding them as an individual.

Table 5 *Test Interviews*

Type of Company	ID code	Position	Project Domain	Background	Interview length (hr:min)	Language	Gender
Private Company	PC0	Data scientist	Consultancy	Computer Science	00:55 + 00:25	EN	M
Academic Research	AR0	Lecturer	Library	Design Informatics	00:34	EN	M

4.3 Ethics

All informants received an information letter and consent form in advance of their interview. They were informed about the purpose of the project, their rights to data privacy and withdrawal from the study. The project, data storage, information letter, and consent form were approved by the Norwegian Centre for Research Data (NSD), prior to execution. Participants were given the opportunity to give consent to individual parts of the permissions asked for if there were items they wanted to opt out of, such as recording. Participants were asked for consent twice because additional permissions were required to gather particularly sensitive information. See the consent forms in Appendix D and Appendix E .

Data was stored according to the storage guidelines of the University of Oslo. Informants' names were de-identified and replaced with ID codes in the data and in this thesis. The keys connecting the ID codes with participants were stored separately from the data. The collection of sensitive personal data related to sexual orientation, race, etc. was limited by asking the number of marginalized identities (see Results, chapter 5.1) rather than their specific orientation, race, etc.

Because the field of AI in Norway is small, it is important that the informants' privacy are protected. Informants were therefore given the choice on how their information were presented in the findings chapter. A tone of moral superiority were strived to be avoided when presenting the informants perspectives and practices (Taylor et al., 2016).

5 Findings

5.1 Results

Table 6
Table of Informants

Type of company	ID code	Position	Project domain	Background	Interview length (hr:min)	Language	Gender
Private Company	PC1	Doctor and researcher	AI in endoscopy	Medicine	00:28	NO	F
	PC2	Data scientist	Finance	Computer science	1:02	EN	F
	PC3	Manager Data science, AI/ML	Insurance	Mathematics	1:02	EN	M
	PC4	Data scientist	Consultancy	Molecular biology	00:57	EN	F
Startup	SU1	Top management	Democratization of AI	Computer science	00:41	EN	M
	SU2	Product manager	Document retrieval	Computer science	00:59	EN	M
	SU3	Top management	Scientific text and data	Entrepreneurship	00:55	EN	Other: Non-binary

Academic Research	AR1	Researcher	Biological neurons	Biotechnology	00:48	EN	-
	AR2	Professor	Informatics	Computer science	00:45	NO	M
	AR3	Researcher	Explainable AI	Physics	00:54	NO	F
	AR4	Researcher	Human Computer Interaction	Psychology	00:34	EN	M
Government Institution	GI1	UX design manager	Pre-trained AI models for image-to-text and speech-to-text.	User experience	00:53	NO	M
	GI2	Analyst	Data analysis	Natural sciences	1:00	EN	M

Note. Abbreviations of participant codes are denoted by the sector they work in: Private Company (PC), Academic Research (AR), Startup (SU), Government Institution (GI). Languages are abbreviated English (EN), Norwegian (NO). Genders are abbreviated Female (F), Male (M).

Table 7
Table of Aggregated Interview Metadata

	Official interviews	**Test interviews**
Gender	Male: 7 Female: 4 Non-binary: 1 No data: 1	Male: 2 Female: 0 Non-binary: 0
Number of participants divided by sector	Private Company: 4 Academic Research: 4 Startup: 3	Private Company: 1 Academic Research: 1 Startup: 0

	Government Institution: 2	Government Institution: 0
Language used during interview	English: 9 Norwegian: 4	English: 2 Norwegian: 0
Sum of length of interviews	10 hours, 58 minutes	1 hour, 54 minutes
Average length of interviews	51 minutes	-
Total number of participants	13	2

Note. Statistics for follow-up interviews are not included in the table.

5.1.1 Interviews

More than 20 people were emailed and asked to participate in this research project. 13 interviews were conducted. The 13 interviews were conducted during a 3-month period from September to November 2020. See Table 6 for information about the participants. The table includes their pseudonym codes and information about their genders, current job position, what kind of AI project domain they work in or have been interviewed about, their educational or career background, the length of interviews, and which language the interview was conducted in.

Some of the informants' answers were edited out because they wished to revise their statements mid-interview to protect the confidentiality of unpublished research and company rules. After a thesis draft was finished, users were given the opportunity to correct the meaning and interpretation of statements in order to validate the interpretation of statements (Taylor et al., 2016). Four out of thirteen informants used this opportunity to provide corrections and clarifications.

Follow-up Interviews

Initially, not everyone was asked about causes and solutions because the assumption was that those who expressed a lower level of knowledge on the issue might not have the best answers to the solution. They were instead asked about how to motivate the organization they belonged to so that it would start working on issues of gender bias in AI. However, this data gap complicated the analysis so the gaps were instead filled.

Follow-up interviews were conducted over the phone in order to fill any gaps in the data that resulted from time constraints in the main interview. These interviews lasted 5-10 minutes and were conducted from April-May 2021. The interviews asked questions regarding their background information in Table 6, their genders, and their number of *marginalized identities*. See below for definition on marginalized identities. In addition, some participants were asked to answer previously omitted questions from the interview guide. 12 out of 13 informants took part in the follow-up interviews.

Marginalized Identities

In this thesis, the term *marginalized identities* refers to minority identities that are subject to and at risk of social discrimination. Some examples of such identities are minority or marginalized identities related to race, gender, mental and physical disabilities, age, sexual orientation, and class. For instance, "white" is a racial identity but since whiteness is not subject to systemic racism, it is not counted as a marginalized identity. Instead, all other races that *aren't* white are considered marginalized identities.

Gender was collected in order to document the gender balance among the participants and to see whether women knew more about the issue. However, gender might not be the only factor that affects how much knowledge an informant might have on the issue of bias in AI since there are several types of biases, as outlined in chapter 2.2. There are other marginalized identities that might increase a person's interest to solve issues of unfair AI systems. Furthermore, without an intersectional view where factors such as race is considered, the full picture might be distorted and one risks accommodating for white women only (S. M. West et al., 2019). In order to assess whether diversity was related to different understandings, information on the participants' subjective experiences of identities was collected.

The question in the follow-up focuses on the informants' subjective experiences of what they identify with and does not distinguish between whether their reported identity is a subjective experience or a scientifically defined category. For instance, there are not necessary any hard limits on what age a person is likely to experience ageism. In a book chapter of *Aging, Ageism and Abuse*, no clearly defined age is found and the chapter cites research related to both people over 60 and people over 65 years (Brownell, 2010). The interview guide for the follow-up interviews are *not* included in appendix to increase the protection of the informants' anonymity. For the same reason, Figure 5 lists the numbers groupwise for those who have marginalized identities as they might belong to an easy identifiable minority.

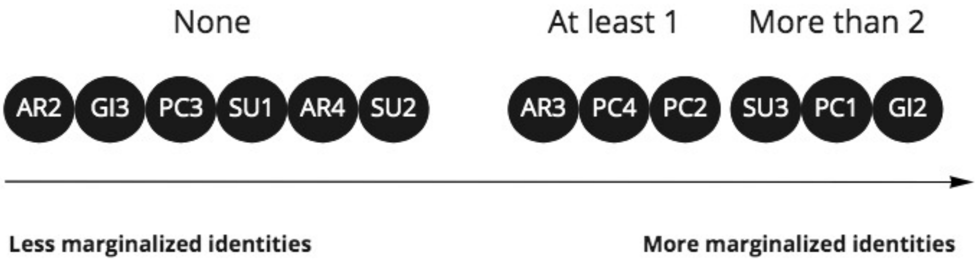


Figure 5. The informants' number of marginalized identities. The informants are groupwise ranged from zero to several marginalized identities. Marginalized identities refers to factors such as race, gender, disabilities, age, class, sexual orientation, and other social barriers. Individual numbers of identities are omitted to increase the anonymity of informants. AR1 is not listed because no data on identity was collected.

Pros and Cons of Digital and Physical Interviews

Due to the pandemic COVID-19 some of the interviews had to be conducted over Zoom video meetings. The benefit of virtual interviews was that I could test interview participants in other cities and that the logistics of the meetings were easier. Additionally, since people could be in their own homes and offices for the virtual interview, it might have increased their level of comfort (Taylor et al., 2016).

The drawback of virtual interviews was that the body language of the participants and the interviewer had limited visibility. Although this thesis does a content analysis of the interviews so the body language is less relevant for the analysis, the lower visibility of body language might have made it more challenging to make the participants feel safe enough to speak freely (Robson, 2002). The topic of gender discrimination is sensitive and some participants might feel a pressure to appear a certain way.

An additional benefit of physical interviews is that because it is easier to read each other's reactions, this could allow for a more relaxed tone. This informality could be related to why some interviews became longer and more candid for some informants.

5.1.2 Questions That All Participants Were Asked

Below is the essence of the questions that all participants were asked. As this was a semi-structured interview, the phrasing might have varied for different participants. See Appendix B for full interview guide.

The strategy was to start the interview with open and general questions in order to see whether they would mention gender bias as one of the ethical issues without being prompted. When bringing up the issue of bias or gender bias they would be given different examples depending on their domain of work. For instance, for someone working with Natural Language Processing the example could be about how Amazon's resume filtering AI would discriminate against women based on the phrasing on their resumes.

1. What is your experience working with AI?
2. Are there any ethical concerns you need to address in your work and can you elaborate on that?
3. Has the issue of bias been discussed in your organization?
4. Has gender bias been discussed as one of the biases in the organization?
5. How or in what ways is gender bias taken into account in your work?
6. What do you think might be the cause of gender bias in AI?
7. Several organizations and companies have published ethical guidelines for developing AI systems. Does your company use one and why or why not?

As the interviews progressed, it became more apparent that most companies do not use ethical guidelines for AI and that simply asking the previous question might not yield very useful information. I was then inspired by the anecdotal example by Grady & Wallston, where the researcher changed her perspective on the research topic by changing the question to a positive one: "What would encourage women to do breast self-exams rather than why don't they?" (as cited in Maxwell, 2012, p. 46). The following question was included for the 10 subsequent interviews:

8. What do you think would be a good way to motivate your organization to try to solve issues of gender bias in AI?

5.2 Findings From Thematic Analysis

5.2.1 Phase 1 – Transcription

Interviews that were recorded were transcribed non-verbatim for content analysis. Interviews were roughly transcribed using the voice to text function on Google Docs and then proof-read to manually "clean" the text. Interview subject PC2 did not give their permission to be recorded, so for that interview there were only interview notes available.

5.2.2 Phases 2 & 3 – Generating Codes

The coding of the themes was conducted using the software Nvivo 12 (NVivo, 2021). A mix of deductive and inductive coding was done. The focus was to deductively code patterns related to the theoretical philosophies and perspectives on technology. A coding table in paper was used during coding to get an overview of which semantic and latent themes had already been coded.

Braun & Clarke define semantic themes as themes that simply describes *what* has explicitly been said, whereas latent themes interpret and infer less obvious concepts based on what was said (Braun et al., 2018). Semantic themes were identified to get an overview of the participants' answers to specific questions. Latent themes about their beliefs and perspectives were gathered through interpretation of the data. Some inductive coding was done to stay close to the data.

The previously mentioned paper coding table included:

- Which interview-guide questions they had been asked.
- Semantic themes such as *bias*, *fairness*, *accountability*, *transparency*, *explainability.*
- Latent themes such as *their motivation, who they think has power, what they optimize for in their work, separation of values and technology, technical heritage, equity, objectivity*, and *intersectionality*. For full table, see Appendix F .

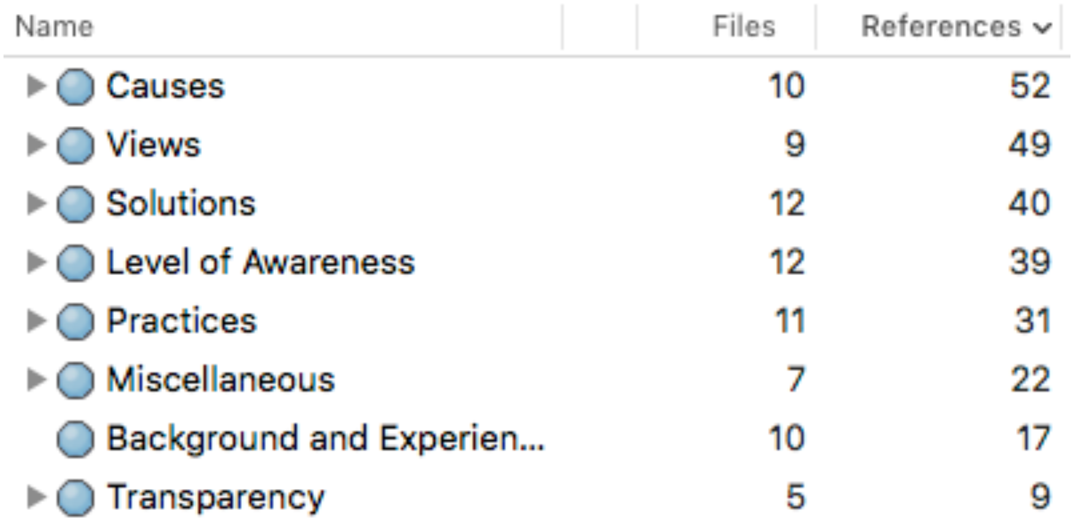
Approximately 389 themes were coded for the dataset in order to make sure that nuances of similar answers were included. One snippet of text could have more than one code. Figure 6 and Figure 7 show an example of how the coding was done in Nvivo 12. The themes were then put into several levels of overarching categories of themes. The 8 very top levels that were defined were *Causes* to GBAI, *Views* that they had on technology, *Solutions* they suggested, their *Level of*

Awareness on GBAI, ethical *Practices* that they had, their *Background and Experience*, *Transparency,* and *Miscellaneous*. See the top level themes in Figure 7.

Table 8
Example of the Coding Process - Phases 2 and 3 in the Thematic Analysis

Direct quotes	Semantic themes - Descriptive	Latent themes - Interpretive
'I realised there was this kind of two-sided side to AI. Most projects use AI from a top-down perspective where companies use AI to analyse private persons, but there was very little in the opposite order, there was very little AI that was used by individuals to analyse corporations or sort of larger entities. And so I was kind of thinking about this power balance issue.'	• Background and experience • Unbalanced power in AI seen as part of the problem	• The power is with the big tech companies • Considers power structures of society

Figure 6. Screenshot of the transcript of SU1 in Nvivo 12 with coding stripes to the right. The coding stripes indicate at which lines in the text a certain theme has been coded.



Name	Files	References
Causes	10	52
Views	9	49
Solutions	12	40
Level of Awareness	12	39
Practices	11	31
Miscellaneous	7	22
Background and Experien...	10	17
Transparency	5	9

Figure 7. The top level themes of coding in Nvivo 12. The number of references show the total number of times the sub-level themes within have been coded. The number of files show how many files are referenced within that theme.

5.2.3 Phases 4 & 5 – Reviewing and defining themes

The software XMind (XMind, 2021) was used to create thematic maps to gain overview of the coding as suggested by Braun and Clarke (2006). Figure 8 and Figure 9 show these thematic maps and they contain about 40 and 66 themes, respectively.

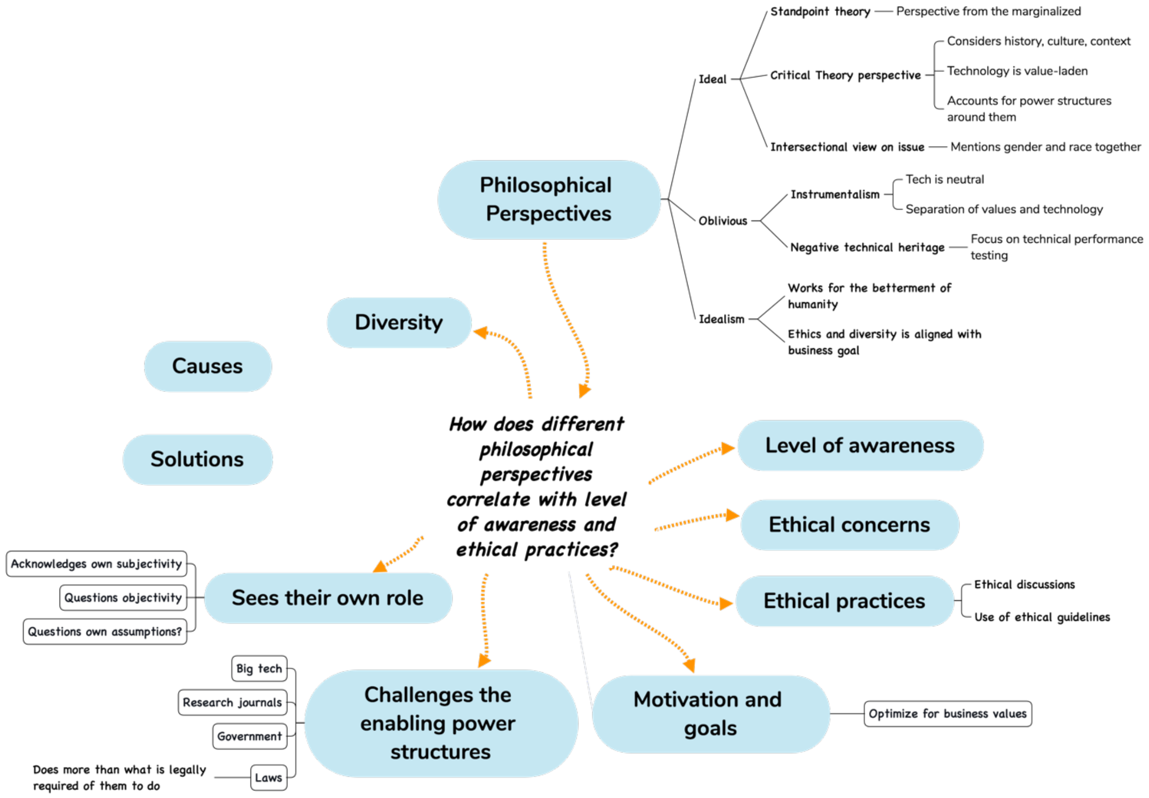

Standpoint theory
Perspective from the marginalized
Considers history, culture, context
Ideal
Critical Theory perspective
Technology is value-laden
Accounts for power structures around them
Intersectional view on issue
Mentions gender and race together
Philosophical Perspectives
Instrumentalism
Tech is neutral
Separation of values and technology
Oblivious
Negative technical heritage
Focus on technical performance testing
Works for the betterment of humanity
Idealism
Ethics and diversity is aligned with business goal
Diversity
Causes
Solutions
How does different philosophical perspectives correlate with level of awareness and ethical practices?
Level of awareness
Ethical concerns
Acknowledges own subjectivity
Questions objectivity
Questions own assumptions?
Sees their own role
Ethical practices
Ethical discussions
Use of ethical guidelines
Big tech
Research journals
Government
Does more than what is legally required of them to do
Laws
Challenges the enabling power structures
Motivation and goals
Optimize for business values

Figure 8. A thematic map of the thematic analysis during phases 4 & 5.

Figure 9. A more detailed thematic map of the thematic analysis during phases 4 & 5.

Tables with extracts of transcripts were made to get an overview of the participants' coding. The table included each participants' perspectives on technology, ethical concerns, awareness around gender bias in AI, how they viewed responsibility, their ethics practices, development practices, and what they viewed to be the causes and solutions to gender bias in AI. Each case was independently analyzed and then compared with other cases. These tables are not available for preview due to the protection of the informants' anonymity.

A second round of analysis was done to find higher level interpretive themes based on the themes from the tables. These final themes are outlined in the next sub-chapter below. Because the thematic map is too complex and big to fit an A4 page, it can instead be found online here: https://miro.com/app/board/o9J_ktfvCzk=/ (password: genderbias)

5.2.4 Overview of Findings and Themes

After conducting a thematic analysis 10 main themes were found, see Table 9. The themes resulting from the thematic analysis are listed below, organized around the research questions. They will be further outlined in Sections 5.3, 5.4, and 5.5.

Norwegian quotes have been translated to English to the best of the researcher's ability. Other quotes are edited for clarity without changing the meaning of the content. E.g. removing repetitive words or superfluous words mainly used in verbal communication such as "like" or "you know", or removing mental side-tracks that are not related to the allotted theme. The symbols *(...)* denotes that some irrelevant text in between has been removed. [pause] denotes a brief pause in the conversation. The researcher's clarifications on context or notes are also in brackets []. Emphasis denoted in cursive are the informants' own emphasis. Since PC2 did not allow recording of the interview, quotes are from research notes written during the interview. The notes attempted to stay close to PC2's phrasing.

Because this is a qualitative study, frequency counts are not the focus of the study, and some of the abstract concepts are not easily quantified. Some places will list a specific count where possible, whereas, in other places where only an indication of frequency is relevant, numbers are substituted by words. In this thesis, *some* refers to 1-3 informants, *several* refers to 4-7 informants, while *most* refers to 8-13 informants.

A draft of the thesis was sent to all the informants prior to submission so that the informants were able to see how they were quoted. Changes in the thesis and omissions of quotes were made to ensure that the findings were true to the informants and reflected what they had meant. According to Taylor et al. (2016), to have informants review drafts increases the quality of the study.

Sub-question 1:
What understandings of technology are found among AI practitioners?

Summary of Findings

The informants' understanding of technology was affected by the technical heritage of the field of computer science and AI. Most of the informants showed traits of instrumentalism and separated themselves from the technology they create. Several of the informants showed a varying degrees of critical perspectives on how practices are done or have been done, and they considered technology in relation to themselves, society, history, power, and other people. Most of the participants had instrumentalist perspectives and several informants also had some critical perspectives.

Sub-question 2:
How does gender bias come into an AI system?

Summary of Findings

The informants had varying degrees of awareness and knowledge on GBAI. The informants reported that gender bias comes into an AI system through causes like biased data, human bias, lack of diversity, and the lack of a definition of fairness.

Sub-question 3:
What practices are in place to detect and address gender bias in AI?

Summary of Findings

The main practices that were found were the use of ethics guidelines, testing for biases, balancing for gender in datasets or data collection, and increasing diversity.

Table 9
Overview of Themes

Research Question	Theme
1 – Understandings of technology	**Perspectives affected by technical heritage**
	Perspectives with traits from Instrumentalism
	Critical Perspectives: Technology is considered in relation to themselves, society, or history
2 – How gender bias enters AI systems	**Causes related to data: Data bias, biased data collection, human bias**
	Other causes: Lack of diversity, missing definition of fairness
3 **–** Implemented Practices	**Formal and informal regulations: Ethics guidelines and laws**

	Testing for biases
	Gender balance data
	Increase diversity
	Delegation of responsibility

5.3 Sub-question 1: What Understandings of Technology are Found Among AI Practitioners?

5.3.1 Perspectives Affected by Technical Heritage

All of the participants seemed to be affected by the technical heritage of AI to some degree, but some more than others. As explained in chapter 3.1.2, Feenberg defines technical heritage as the assumptions, practices, and perspectives on the world that have been inherited from previous people within a field (Feng & Feenberg, 2008).

There were different aspects of technical heritage observed in the participants. One of the more important ones was that several of the informants only talked about testing for performance when testing the AI. Some of the informants saw AI as a tool that could solve many of our social problems such as inequality and discrimination.

Some of the informants had a very technical view of their work. For instance, PC4 referred to biases as "biometrics not well trained" rather than referring to it as racism. AR3 informant refused to use the word "race" and instead called it "ethnicity".

A perspective affected by technical heritage was to view something that always had been done a certain way before, legitimized the practice today. PC3 said that because their industry had always been using algorithms to calculate different prices for different customers, this ethical issue would be no different with AI implemented: "This is how we have always done things before AI, so it is therefore also fine when using AI.".

Table 10
Perspectives Affected By Technical Heritage

Code number/ Codes		Example quote	Participants
1	**Technical view. E.g. refers to biases in a technical manner**	"(. …) data samples that are collected are just, they're like, very focused on one particular type of person or one particular nationality (. …)" [AR1] When asked about the causes of gender bias in AI: "I don't know that because [pause] I guess it's some biometrics that is not well trained." [PC4]	2 PC4, AR1
2	**Focused on technical performance only when asked about testing**	"When is it good enough? I mean you can always continue updating it (. …). Eventually you're going to reach a point where the performance will be tapering off, so you can spend a lot more time, but you won't be able to do much more. (. …) it is in the greatest sense more related to how good your dataset is, and if you have enough data for training. (...) And then there's only so much you can do to get a great performing AI. (…)" [SU1]	5 AR2, SU1, GI1, AR4, SU2
3	**Technological solutionism**	"I don't think machine learning can solve all of our problems, but discrimination are kind of one of those things that would have been really easy to solve with machine learning – If we reach a consensus about it and it was legally required." [AR3]	2 SU2, AR3

Note. The column *Participants* lists the total number of participants with that perspective and their respective ID codes.

5.3.2 Perspectives With Traits from Instrumentalism

Several of the informants had perspectives on technology that had traits from Instrumentalism. See table Table 11 for detailed codes, quotes, and counts.

One trait was that some of them assumed that some technical aspects did not contain biases. However, at least one of these aspects, word embeddings, has previously been shown to recreate biases (Bolukbasi et al., 2016). This is an indication that they believe technology to be neutral.

Another trait was that some of the informants separated human values from technology. Some of them perceived the data as biased whereas the algorithm is neutral, and some of them attributed biases to the humans in the loop of the AI and not to the AI itself. Several informants also expressed themselves in a way where they were distancing themselves from the issue. For instance, AR2 said that they try "to remove ourselves from the ethical". AR1 knew about the issue of GBAI but did not participate that much in the ethics activities of their organization.

Table 11
Perspectives With Traits From Instrumentalism

Code number/ Codes		Example quote	Participants
4	**Separated human values from technology**	"(. ...) we try, strictly speaking, to remove ourselves from the ethical, but to of course have it in the back of our heads. (...) Sometimes we are so far away from the humans in the study that we think that this is just data that we train on." [AR2]	5 AR2, SU1, GI1, SU2, SU3
5	**Assumes that some technical aspects do not contain biases**	"Right now, our researchers are working on word embeddings models and figuring out how to separate between the word table and word table in different, like, it's not the same. It's not the same dangers in what we do." [SU3]	2 SU1, SU3

Note. The column *Participants* lists the total number of participants with that perspective and their respective ID codes.

5.3.3 Critical Perspectives

Several of the informants showed varying degrees of some level of critical thinking towards technology and the power imbalances in AI, by considered technology in relation to themselves, society, history, and other people.

Some of them considered their own role and subjectivity in relation to what they make. Some wanted to democratize AI and was developing a platform where datasets were to be merged and where different types of people could look through the datasets and say something if they were underrepresented in the dataset (SU1). Some sought the opinion of those who were blind in order to test an AI for generating alternative text (GI1).

Some of them questioned the dominant ways of thinking about gender and objectivity. They referred to gender as a spectrum rather than binary or questioned their own objectivity and the objectivity of research (G12; SU3).

Some used technology to challenge power structures that they saw as contributing to biases (SU3; SU1). For instance, one informant circumvented the citation system when making an AI recommendation system for research because they believed that to be a biased system (SU3). They said that the system is affected by "who knows who?" and who has the power of admitting accepting conference papers.

Several of the informants considered the role of history, culture, or context when talking about biased data and the development of AI. E.g. One informant said that it does not make sense to talk about fairness without context (PC2).

Several of the informants considered how power affected the field of technology. Some of them found it worrisome that there is a skewed power balance between the big tech companies and the people who use the services of big tech.

Some of the participants talked extensively about that it is unknown how to define fairness in a way that is programmable. They said because it is mathematically impossible to not discriminate against one group or another when trying to balance datasets and systems, the issue becomes political. AR3 also saw the issue as political because the reason that there is no institution for algorithmic audits in Norway is that neither of the politicians wants to be the first to suggest this.

Table 12
Critical Perspectives: Technology is Considered in Relation to Themselves, Society, or History

Code number/ Codes		Example quote	Participants
6	**Considered history, culture, and/or context**	"It is kind of not the technologists alone who can sit and develop AI for the health institutions. It has to be put in a context by doctors and other clinicians to understand, meaning to set the preqreuisites, for the development of it." [PC1]	5 PC3, PC1, PC2, AR3, GI2
7	**Accounted for power structures in**	"(...) besides all the privacy issues (...) we often meet people at the most vulnerable. (...) Our users	5

	society	are not users that, you know, are voluntarily with us". [GI2]	SU1,SU3, PC2, AR3, GI2
8	**Saw the issue as a political problem**		2 PC3, AR3
9	**Worried about the skewed power imbalance in AI between big tech and the people**		3 SU1, PC2, AR3
10	**It is mathematically impossible to not discriminate one group or another**	"(. ...) whenever you try to do personalization, you will get into this conflict. (. ...) what groups shouldn't be discriminated? Because it's hard to not discriminate any groups, right? That's mathematically impossible, more or less(. ...) And this usually comes into sort of a political debate, right?" [PC3]	2 PC3, AR3
11	**Considered the perspective of the marginalized**	"(. ...) what's the, sort of the, the contextual rig of this solution? (...) What is it that this solution tries to do? And how does that solution impact the user from a user perspective?" [GI2]	3 GI1, SU1, GI2
12	**Questioned their own objectivity or the objectivity of research**	When asked whether GBAI could be an issue in their AI system: "Not that I have spotted? Again, that's sort of, in some ways, the beauty of scientific research it is attempting, at least, to be objective. Now I have a problem with the fundamental notion that science is objective, because I don't think it is, I don't think it ever is." [SU3]	2 PC1, SU3
13	**Gender is seen as a spectrum and not as a binary**	" Whereas the men and women [pause] You know, at least that binary understanding of gender, the [pause] it's very, very, very often you'd expect them to be [pause] the distributions to be equal." [GI2]	2 GI2, SU3

Note. The column *Participants* lists the total number of participants with that perspective and their respective ID codes.

5.4 Sub-question 2: How Does Gender Bias Enter an AI system?

Summary: The informants reported that gender bias comes into an AI system through causes like biased data, human bias, lack of diversity, and missing a definition of fairness. Another cause is the recurring theme of their forwarding of responsibility.

An overarching theme of several causes and solutions mentioned was to increase awareness and knowledge. These practices were (listed in descending order from the most suggested to the least): Increase practitioners' awareness; awareness in the population; to test for biases; interdisciplinary teams that work closely with a domain expert of the relevant field where the AI will be used; that more companies should know that good practices ethics means good business; universal and international ethical guidelines; a dedicated ethics person that one could ask questions; concretely defining what fairness means; and knowledge-sharing among practitioners.

Most of the practitioners pointed to biased data as the main cause of GBAI. Several pointed to human bias and biased data collection as contributors to biased data. Some of the informants suggested therefore to balance datasets as a solution.

Several of the practitioners said that lack of diversity was a big cause of GBAI:

> *(. …) gender bias, specifically, if we'd start with the most obvious is that and I don't have the statistics of this, but a lot of AI developers are male. (. …) I don't think there's any ill intent (. …) there is nothing wrong with being a white guy in his mid 20s. (…) But if you get a room full of people who all have the same lived experience, well, they're not going to question and think about other lived experiences. (. …) if you'd had a roomful of, you know, black 70 year old women to developing these tools, you would be a little bit of a whole other set of biases into it (. …) you have to look at every challenge you have from different angles to understand (. …) [SU3]*

Some of the informants saw increased diversity as a way to increase perspectives and therefore avoid blind spots and biases. Several of the practitioners suggested therefore to increase gender diversity to address the issue. One informant said that lack of diversity is probably the cause because men are less interested in solving issues like GBAI (SU2).

Some said the cause was slow regulations and several informants suggested that changing regulations as a solution. Some pointed to the lack of a clear definition of what fairness entails as the cause of GBAI and the lack of regulation. Some said that increased accountability is needed and several suggested that algorithmic audits should be implemented. However, one of those who suggested algorithmic audits said that voluntary knowledge-sharing among AI practitioners would be better than strict audits.

AR4 believed that there is enough regulation in Norway to safeguard against biased designs made by men. He thought that companies are allowed to differentiate on gender, but not where they live. PC3 who worked with these topics and knew these rules well said that companies are *not* allowed to differentiate on gender, but they are allowed to differentiate on geographical areas.

Some pointed to the structures of society as the cause of GBAI: "AIs that are based on human data or social data is a reflection of our society. " [PC2]

Table 13
Table of Reported Causes

Code number/ Code	Participants who said this was a cause
14 Biased data	11 AR2, PC4, SU1, PC3, AR4, AR1, SU2, PC1, PC2, AR3, GI2
15 Human bias	6 AR1, SU1, GI1, SU2, PC2, GI2
16 Biased data collection	5 AR2, AR1, PC1, PC2, GI2
17 Lack of diversity	5 GI1, AR4, SU3, SU2, PC2
18 Need definition of fairness	4 PC3, AR3, PC2, GI2
19 Structures of society	3 AR3, PC2, GI2
20 Slow regulations	2 SU1, SU3

5.4.1 Delegation of Responsibility

One of the causes and themes found among the informants was their delegation of responsibility. The informants who were in a position to teach taught their students very little or no AI ethics. One informant forwarded the job of teaching students about AI ethics to the ethics specialists who taught general courses on ethics.

Some of the informants forwarded the responsibility of making the companies change their gender-biased algorithms onto consumers. Some of the informants said that one of the causes of gender discrimination is that consumers want personalized online experiences. Some of the informants suggested that increased transparency of companies and awareness in customers would lead to the market regulating itself because customers could demand AI systems that did not discriminate. Another informant, on the other hand, meant that market regulation does not work very well.

> *There is this other [pause] "kind of" [informant made quotation marks with hands] regulating mechanism that we have in market capitalism, which is reputation. That the customers don't like them anymore if you discriminate. But we see that self-regulation in a capitalistic environment works really shitty. [AR3]*

Some of them forwarded the responsibility of biased data onto the data providers; the customers or those who made the datasets. Some of them assumed that the datasets they used were without bias and did not check for biases in them.

One of the informants forwarded the responsibility of biased benchmarking standards onto the research institutions who made the datasets and onto the big tech companies who had the resources to train large neural networks. One informant used an image dataset from MIT called TIMIT that he knew were biased to benchmark image recognition AIs, despite his awareness of cases of GBAI (AR2). He said that it is necessary to use the "gold standards" for benchmarking if one wants to be published. He said that making a new standard for benchmarking would be easy, but that it would be difficult and require a lot of lobbying to convince the rest of the research community to use the new standards.

Furthermore, the informant rarely trained an algorithm from scratch despite knowing the pre-trained ones might be biased (AR2). The bigger pre-trained AI algorithms from Google had a higher performance than any algorithm he could make with limited data resources, so he usually just trained the last part of the neural network to fit his projects.

> *If I had trained everything from scratch, I probably would have both performed worse on the gold standard, because it is already trained so much, but I would also in practice be wasting resources because I would not be solving the problem I am trying to solve any better. (...) Someone would then have to train everything all over with a properly balanced dataset. But there is nobody who does that. It costs a lot.*

5.5 Sub-question 3: What Practices Are in Place to Detect and Address Gender Bias in AI?

The main practices that were found were the use of ethics guidelines, testing for biases, balancing for gender in datasets or data collection, and increasing diversity. There were additional practices that were implemented by GI2's organization, but they would potentially be identifying GI2 if they were listed in detail. Additionally, the aim of the multiple case studies are that cases replicate and support each other's findings. The Findings chapter will instead focus on the practices that were most commonly mentioned by the informants and the literature.

Some of the informants worked to increase the explainability of AI algorithms. One of the informants had access to ethics people in their organization but they were not consulted for advice on ethics in other research projects. The task of the ethics people was to conduct research on whether AI systems are able to learn morality and to teach courses in ethics.

Several of the informants reported that GBAI had been discussed in their organization. Most of the informants had either not discussed the issue of GBAI in their organization, or they had discussed it, but the topic was limited to whether to gender AI assistants.

4 of the informants used ethics guidelines for AI. Most of the participants either said that they did not use ethics guidelines or they did not mention using any when asked about what their ethics practices were. Several informants asked for a clarification on what "ethics guidelines"

entailed when asked about it. Some replied that they did the ethics practices that were legally required to do, such as GDPR.

One informant had made ethics guidelines for another organization but did not use any in his main organization. Some of the startup informants that did not use ethics guidelines expressed that it is more important to have actual processes for dealing with ethical issues. They said that guidelines could be used for "ethics washing" to improve the appearance of an organization.

Some of the informants had implemented testing practices and were developing control points and processes for bias testing. Some of the informants emphasized that the only way to detect gender biases is to test for them. Some other informants had practices for achieving gender balance in their data and attributed this to the relevance of gender in the medical field and biology, to which they belonged.

Some of the organizations of the informants were in the process of implementing processes and activities for AI ethics. PC2's organization was in the process of implementing an AI ethics department and had held a conference on the topic. The same organization had also pledged to reach some of the Sustainable Development Goals (SDG), such as the SDG on increased diversity. AR1's organization had recently hired a dedicated ethics person and had held meetings to discuss which biases that were relevant for their research projects.

Some of the informants practiced involving users in the AI development. E.g. One organization practiced having multi-stakeholder discussions where representatives of different user segments were included (GI2).

Several of the informants did work to promote awareness of the ethical issues of AI. However, most of the informants had no or very few practices for specifically addressing GBAI. Some of the informants had practices that were adjacent to ethics practices to address GBAI, but they were mainly done for business reasons. Such practices were conducting security risk analyses and discussing how algorithmic price changes affected the customer experience.

One startup took it upon themselves to not perpetuate the biases of citation systems in the research community in their AI recommendation system (SU3). They did this by not using the number of citations as data points and hired more researchers to find other ways of determining the relevance and quality of research data. Similarly, another startup decided to not create personalized search results for their users to avoid creating filter bubbles (SU2).

5.5.1 The Use of Regulations

One informant said that using ethics guidelines for AI is not enough because they are too vague and have too much room for interpretation:

> *(. …) ethics guidelines, they are exactly what they are, it is "should" and it will always have room for interpretation and [pause] context-dependent considerations and personal values (. …). I think that for AI, something stronger is needed. (. ...) we can't "should" our way to a better world. Because AI boils down to deterministic programs, right? I think we need rules. And there, the position of democracy is much weaker. Dictatorships have greater progress on AI because they can just cut through and say this is how the road will be. [AR3]*

AR4, SU1 did not mention whether they did or did not use ethical guidelines, but it did not come up when they were asked about ethical concerns in their work or when asked about how gender bias is taken into account in their work.

Some informants talked about how ethics guidelines can be used for *ethics washing*, companies talking about ethics without actually practicing it (SU3; GI2):

> *(. …) you had like corporate social responsibility being hyped in the, early 2000s, where every company was donating money to a village in Africa, while spewing out local pesticides and killing off the environment. It's ethics washing in a way. And I'm not saying all companies do this. (…) I think a demonstration of principles in the company culture is so much more important than any guidelines you could make. [SU3]*

One informant had contributed to the making of ethical guidelines in a different organization, but did not use any guidelines in his main organization. He mainly did what was legally required (AR2):

> *And there [note: "there" refers to a secondary organization], we develop something called data guidelines, which is sort of ethical guidelines. (…) it is about a lot of things. But discrimination is one of them, that we should avoid that type of thing. But I don't actually know of anything that is in use and practiced, at least not with us [note: "us" refers to current main organization]. It's not like we sit and follow some ethical guidelines outside of what is given, legally given, from what I understand. (…) Not in any systematic way of following it, at least. [AR2]*

Table 14
Practices Related to Guidelines and Laws

Code number/ Codes		Example quote	Participants who said did the practice	Participants who said they did *not* do the practice
21	**Used ethics guidelines for AI**	"We've looked at them and they have inspired the principles that we have. So I think that our principles are in line with what is out there and what I've seen." [GI2] "I have looked at one, but I haven't(...) It's not something that we have sort of done a lot of work around." [PC3]	4 PC1, PC2, AR3, GI2	7 AR2, PC4 GI1, PC3 AR1, SU2, SU3
22	**Mainly did practices that was legally required**	When PC3 was asked whether they monitor whether they discriminate women: "(...) So, so you could, you could put it in place regulation that should(...) 'Okay, we have to make sure that the average between these groups and these groups should be the same', right? But, but we don't. We wouldn't, I don't think we wouldn't mind if that kind of regulation came into place. But since it's not in place, we don't." [PC3]	3 AR2, GI1, PC3	-

Note. The column *Participants* lists the total number of participants with that perspective and their respective ID codes.

5.5.2 Discussions on Gender Bias in AI

Table 15
Practices Related to Discussions About Gender Bias in AI

Code number/ Codes		Example quote	Participants who said did the practice	Participants who said they did *not* do the practice
23	**Had discussed gender bias in AI in their organization**	"Yeah, that is a, that is a major problem. A lot of the data samples that are collected are just, they're like, very focused on one particular type of person or one particular nationality or	7 PC3, SU1, AR1, SU3,	6 AR2, PC4 GI1, SU2, PC1,

		something like that." [AR1] "Not so much, maybe? It might be that we should be talking more about it. I wish the answer was yes, but I think the answers is no." [AR2]	PC2, AR3, GI2	SU3
24	**Had discussed whether to gender AI assistants**	"Yes. chatbots, should the chatbot be a robot, man or female? So should they be gender or not? And if it is not gender *[sic]* or if it is gendered, should it be male or female?" [AR4]	2 AR4, SU3	

Note. The column *Participants* lists the total number of participants with that perspective and their respective ID codes.

23 Had discussed gender bias in AI in their organization

One informant reported that gender bias in their algorithms had been discussed at their organization, however, those discussions had more of a focus on customer experience rather than the social implications of it (PC3).

5.5.3 Testing & Development Practices

Code number/ Code		**Example quote**	**Participants who said they did the practice**	**Participants who said they did *not* do the practice**
25	**Had practices for achieving gender balance in their data**	"(. …) in clinical studies, one has to have a balance, so yes. There is a difference on how you recruit. (. …) there are regular research guidelines that are relevant, that you try to have from both genders. (…) But in the particular development of data, of the AI, *that* is where one has to make a decision: Is it relevant with diversity here? In which case, what diversity?" [PC1]	2 PC4, PC1	
26	**Had test practices to evaluate gender bias**	"Yes, gender biases is definitely one of the issues. It's also one of the things that we're looking at in some of our projects, we can also see that there are [pause] without going into all the details now, that there are gender biases in our processes too (…) I don't think they are completely understood. I don't think they are intentional. And I don't think that, that we are	1 GI2	9 AR2, SU1, PC3, GI1, AR1, AR4 SU2, SU3, PC2

		the only actors in that ecosystem that contribute to those biases, but we do see some others [pause] skewness in some cases, in our data, we do." [GI2]	
27	**Said that bias testing was necessary to not have gender biases in the system**	"So I do, I personally believe in that you could, you know, it's easy to say that you don't discriminate, but I don't really believe it. Because it's so hard not to do it. So when people say, you know, 'our models don't', then show me that, what tests you've done because if you haven't, then it's very likely that it is in some way, right? It's very, very, very hard not to do." [GI2]	2 GI2, AR3
28	**Conducting research on gender differences seen as negative for progress in gender equality**	"I am not very fond of research that compare females and males, and then provide some kind of gender difference. (. …) it's a bit lazy. Because you can always find differences between groups. (…) I think it's the way to reach gender inequalities, what [sic] one way to reach gender inequality is to sort of not spend too much energy on finding the gender differences. (...) And I think that my opinion is maybe stronger on that. (. ...) find some kind of difference and it's interesting because it is from these cultural, ethnic or cultural groups. I disagree. I don't think it is interesting. (. …) gender equality is better achieved by not spending your energy on gender differences. (...) We have been spending our resources researching cultural and ethnical differences for centuries. So we have enough of that, we're sort of, we're done." [AR4]	1 AR4
29	**Was developing control points for bias testing**	"So we *have* spent a lot of time thinking about how to, you know, self-regulate, and also put mechanisms in place that allow us to unpick some of the problems that could arise, they're not always easy to spot. (…) Where should you put your control points, And what should you control for at different points in your pipeline?." [GI2]	2 GI2, PC2

#25 Had practices for achieving gender balance in data

Some informants do projects in the field of medicine and biology and they attributed the gender balancing practices on the nature of the fields (PC1; PC4). One informant expressed the importance of finding the confounding factors in the data, whether a correlation was real or whether something else made it look like a correlation (PC4). One informant emphasized the importance of having gender balanced data in order to draw the right conclusions, because she had seen examples of studies where a gender balance was missing and incorrect conclusions were drawn (PC1). Some of the informants seemed to have some awareness of the need for an intersectional approach because they mentioned the need to assess relevant fairness metrics or types of diversity (GI2; PC1).

#26 Had test practices to evaluate gender bias

One informant said that they were not able to test for gender bias in one project because they were not allowed to collect gender as a datapoint due to privacy issues (AR2). They said this removal might mean the algorithm would perform worse.

When it was mentioned to that voice recognition has previously worked less well for women, he realized that gender had been a blind spot during the testing of AI-systems for auto-captioning videos (GI1). The informant had only taken out random samples and tested for aspects like dialects and whether English were their native language.

5.5.4 Diversity in Genders & Nationalities

All of the informants, except for one (PC1), reported a lack of diversity in their organization and some of them had no female permanent employees. The one informant was the only one with 50% women on the team (PC1). Several of the informants' organizations were therefore trying to increase diversity in the organization. Most of them still had a higher level of diversity either in their department or organization than the average of the IT field in Norway, which was 20% in 2016 (Brombach, 2016).

Several of the organizations had some female or non-cis-male leadership in the department or organization. Some of the informants with more diversity in the organization saw practices for ethics and diversity as strategic business decisions and not as something "nice" to do. The organization of one informant had even committed to United Nations' Sustainable Development Goals (PC2). This perspective might suggest the reason why their organizations had more diversity.

The startup that trained people from other fields to become AI practitioners had a high level of diversity and had no conscious practices to increase diversity (SU3). The hiring and training was done through a service of a company called Science 2 Data Science that helped STEM scientists transfer into the field of AI.

The other startup, was looking for women with technical AI skills had very few female applicants and had not been able to find any women they deemed to be qualified (SU2). Surprisingly, they had deliberate strategies and knowledge on how to recruit women. They avoided certain words that would deter women, something they had learned at a conference for diversity. But they still only got about 1% female applicants to one of the job postings, despite receiving more than several hundred applications. They attribute this challenge to having to compete with the big AI companies and that maybe girls are not as interested in having a startup as a workplace (SU2).

At least 40% of the employees in the two aforementioned startups were of foreign nationalities (SU2; SU3). They attributed the high level of national diversity to having fully distributed teams. They could hire people from the entire world because they did not require that people lived in Norway. The findings suggest that a change in hiring practices might lead to higher levels of diversity in both gender and nationalities.

Table 16
Table of Diversity Statistics and Practices

Code number/ Code		Example quote	Participants who said they did the practice	Participants who said they did *not* do the practice
30	**Diversity: Had a higher percentage of female employees than the average in IT in Norway***	"We had 42% women at some point. And that was including in our tech teams. (...) It does help, I think, to have top level management that are diverse. (...) We've never had a deliberate strategy to get more female [employees] in." [SU3]	Total: 10 50% women in team: 1 PC1 30-40% women in department: 2 GI1, AR3 25-30% women in department: 1 GI2 30-40% women in organization: 6	Total: 3 No female permanent employees: 3 AR2, SU1, SU2

			PC4, PC3, AR4, AR1, SU3, PC2	
31	**Diversity: Had female leadership**	Some of these numbers were confirmed by annual reports from their respective organizations. These reports can't be listed to keep the anonymity of informants.	5 Has at least 30% women in top leadership positions: 3 AR4, PC3, PC2 Leadership in department or organization that includes someone not cis-male: 2 SU3, GI2	
32	**Diversity in nationalities**	"(...) from the very start, we were just a team of people who would who had to figure out how we deal with our diversity, diverse cultural backgrounds etc. (. ...) a lot people have like two or three nationalities in one. " [SU3]	At least 40% of employees were not of Norwegian nationality: SU2, SU3	
33	**Their organization tried to increase gender diversity**	"We try to make sure that we have a gender balance in on the project team. If we have a choice between male or female project leader, we often choose female project leader." [AR4]	7 AR2, PC3, AR4, AR1, SU2, PC2, GI2	
34	**Diversity was seen as a strategic business decision, not something that was "nice" to do**	"(. ...) a misunderstanding of what diversity is, because diversity isn't about being nice. Diversity is a really, really important business decision. Because if you do build biases into your algorithms, first of all, you're building a really shitty world. And second of all, you're painting yourself into a corner that sooner or later, someone's going to hold you up on this and say, 'what the fuck are you doing?' " [SU3]		3 PC4, PC2, SU3

*The average female percentage in Norway in 2016 was 20% (Brombach, 2016).

Causes of Gender Bias in AI and Perspectives

6 Discussion

6.1 RQ1: What Understandings of Technology are Found Among AI practitioners?

6.1.1 Understandings Affected by Technical Heritage, Instrumentalism, and Critical Perspectives

The majority of informants expressed an instrumentalist technology perspective, while some of them combined this perspective with a more value-based perspective. It was clear the technical heritage of the field of AI had an important effect on their understanding of AI as a technology.

The separation of values and technology or research is confirmed in a recent study that analyzed some of the most cited AI conference papers (Birhane, Kalluri, & Card, 2020). The authors found that most AI papers focus on technical values such as performance and accuracy, and that the papers largely do not consider social and human values. They also found that a common attitude among AI researchers was to consider their research as apolitical, which is similar to the attitudes found among several of the informants.

Although the informants with some critical perspectives were affected by the technical heritage as well, they also considered the effects of history, power, society, and their own subjectivity on technology and their work. This perspective seemed to make them less affected by the technical heritage and is more in line with the Critical Theory of Technology (CTOT) that Feenberg proposes. CTOT emphasizes the role of power, history, and the way society is designed in the devices that are created (Feng & Feenberg, 2008).

Feenberg says that the Instrumentalist perspective separates humans from technology and views technology as neutral. He claims that although the technical elements that make up a device can be neutral, biased devices shaped by society and the norms of design should not be mistaken as neutral just because they consist of neutral elements. Here, Feenberg refers to "biased" devices not as devices that discriminate against someone, but rather that the devices hold values and are shaped by the demands of the social world which make them not neutral (Feng & Feenberg, 2008).

The dominant perspective among the informants is that it is the data that is biased but the algorithm itself is neutral. Perhaps because of this phenomenon, most of the informants see the algorithm as a neutral technical element that can be combined with data to become a device, which they believe can *become* biased if the data are biased. In contrast, AI ethics researcher Dignum (2019a) writes that AI is not value-neutral because it embeds the values, interests, and goals of those who create them. A similar position is discussed by AI anthropologist Forsythe, who found that "contested cultural assumptions are routinely inscribed in supposedly neutral technologies" (2001, p. 115).

The algorithm is akin to a statistical filter that controls which parts of the data are emphasized. Although the algorithm changes its analytic behavior and its results after training, most algorithms still inherently look for dominant patterns without questioning them. Choosing which algorithms to use are a non-neutral task when AIs can either be made with algorithms that perpetuate the status quo, or algorithms that actively counter biases. An example of a type of algorithm that actively counters biases during training is Generative Adversarial Networks (GAN). A study found that using GAN would mitigate bias during the training of an AI (Zhang et al., 2018). Other researchers suggest the development of new algorithms that counters the biases embedded in data (Wellner & Rothman, 2020; Zou & Schiebinger, 2018). Perpetuating the status quo is not a neutral action as it preserves current biases (Wachter et al., 2021).

An alternative explanation of why most of the informants view the algorithm to be neutral is found in the work of Verbeek (2008). Verbeek would say that the reason the untrained algorithm, or technology, is seen as neutral is because it does not have intentions or consciousness of its own (2008). The Norman AI has no consciousness and simply makes the associations it is trained to make. However, Verbeek argues that because technology is able to direct our moral course by shaping which decisions we make it has a form of intention and is not neutral; the choices we make are different because of the new choices technology facilitates (2008).

Jones explains the origins of the Instrumentalist approach found in the informants and in the field of AI in his essay on *Data Positivism since World War II* (2018). The author argues that the AI algorithms that are used on massive amounts of data today are not neutral components of a system which simply arose from an "objective" research community that neutrally appointed those algorithms as the best ones. Jones explains the developments in computational statistics and pattern recognition since the 1950s and explains the path to success of today's big data algorithms: A paradigm shift from a realist approach to an instrumentalist approach. The first seed of this approach, he writes, comes from developments in machine learning in the Soviet Union around the 1960s and 1970s. According to Jones, the realists tried to explain the patterns in the data with a focus on scientific truth, but this did not work well on large datasets. The instrumentalist approach focused on finding any function that could represent the data just well enough for predictions. Jones attributes this paradigm shift to the subsequent success of Instrumentalist AI algorithms, such as Support Vector Machines. He adds that the developments in the decades that followed, increasingly gained predictive power and established the value of prediction over explainability (Jones, 2018).

This means that the Instrumentalist approach in the AI field today relies on a historical choice of ignoring how an AI reaches its decision, which includes patterns of sexism and racism in decision-making, as long as it has predictive power. Instrumentalism philosophically ignores whether the algorithms are biased or its societal impact because it separates the means from their ends; it does not matter how the algorithms work or whether they are obscured by patterns of inequality, as long as it leads to the desired result: Accurate prediction.

In business, this can translate to “it does not matter if it is sexist or racist as long as it makes or saves money”. Jones (2018, p. 683) writes: “A generation ago, the inscrutability of neutral nets made them deeply problematic; the renaissance of neural networks from around 2012 rests squarely on the legitimation of such black box algorithms". This could explain Google's decision to mute and fire their ethics researcher Timnit Gebru instead of addressing her findings on how large language models can harm equality and are too large to have any accountability (Jonhson, 2020).

Feenberg says that it is more important to use CTOT to evaluate the technical heritage, such as taken-for-granted assumptions, than to analyze the designer. He says it is not the designers nor their close environment, but their background assumptions which have the biggest effect on the outcome of a design. Feenberg believes that the only way to open a path to different designs where humane values are incorporated is to actively question technology and to seriously consider the effects of technical heritage (Feng & Feenberg, 2008).

More research is therefore needed to map the technical heritage of AI, on how it contributes to GBAI, and which alternative future paths are possible. AI practitioners need to be aware of the inherited assumptions that they take for granted and whether these are assumptions they should keep.

6.2 RQ 2: How Does Gender Bias Enter an AI System?

6.2.1 Causes & Solutions

Most of the causes that the informants mention coincides with existing literature: biased data (Zou & Schiebinger, 2018); biased data collection (Zou & Schiebinger, 2018); lack of diversity (Leavy, 2018; S. M. West et al., 2019, 2019); the need for clear definitions of fairness in order to be programmed and regulated (Dignum, 2019a; Parsheera, 2018; Zou & Schiebinger, 2018).

The solutions give an indication of what the causes are, and most of the suggested solutions are also supported by the literature: increased awareness (Zou & Schiebinger, 2018); interdisciplinary teams (Leavy, 2018; Zou & Schiebinger, 2018); increased transparency (Dignum, 2019b; O'Neil, 2016; Thelwall, 2018; Wellner & Rothman, 2020); universal ethics guidelines (Parsheera, 2018); defining fairness (Parsheera, 2018); algorithmic audits (Raji & Buolamwini, 2019); increased accountability (Dignum, 2019b), align ethics & business goals (Dignum, 2019b); and change regulations (Dignum, 2019a).

However, to view the reported causes and solutions as answers presupposes that the informants are qualified and know the answers to what causes GBAI. On one hand, they had at the very least an education in AI and most of them had had years of experience in the field of AI. They are therefore at the very least deemed to be qualified to have a technical opinion on the causes and solutions to gender bias in AI. If only the answers of those who knew more about GBAI are emphasized, the answers are a bit more systemic: Structures of society, the need for a definition of fairness, and biased data.

On the other hand, as discussed in the previous sub-chapter, their perspectives are affected and limited by the technical heritage from the field of AI, which includes their instrumentalist approaches to AI. Their instrumentalist understanding of technology could explain why the most common causes that they list are the things that are easiest to see from a technical perspective, and rather superficial from a systems thinking perspective, as shown below in Figure 10.

In Figure 10, the causes are organized into Meadows' iceberg model for systems thinking, and in Figure 11, the solutions are organized into her theory about leverage points (LP) as outlined in the theory chapter. The iceberg model is a tool for thinking more systematically and the lower a cause is on the iceberg, the more leverage it has; note that this is the opposite order of leverage compared to the LP model. What most of them see as causes are just the "tip of the iceberg" in Meadows' iceberg model, and with the lowest leverage point; biased data, and biased data collection that leads to biased data. This is a surprise, as these are the causes continuously pointed to in the body of literature assessed for this thesis. Similarly, an often suggested solution within the literature is balancing datasets; this is considered a low LP intervention. As Meadows states, the different causes and solutions can be argued to belong to different levels of impact, so the categorization is not absolute but gives an indication of the differences in severity and impact between the different causes and levers of intervention (Meadows, 1999).

AI anthropologist Forsythe (2001) states that assumptions held by the AI practitioners limit what they are able to see; most of the informants see their work as apolitical and neutral and that might limit them from seeing more systemic causes. Zou and Schiebinger (2018) state biased data as a cause, which is low impact on LP theory, but they also attribute this cause to higher impact systemic causes: "Biases in the data often reflect deep and hidden imbalances in institutional infrastructures and social power relations". Most of the informants did not point to similar systemic reasons when they talked about biased data, and some of them even avoided using words like race and racism. Because of this technical heritage, their answers are seen more as opinions rather than expert knowledge. Some of the informants with a more critical understanding of technology pointed to systemic causes such as the structures of society, which refers to how society is built and impacts people; and slow regulations.

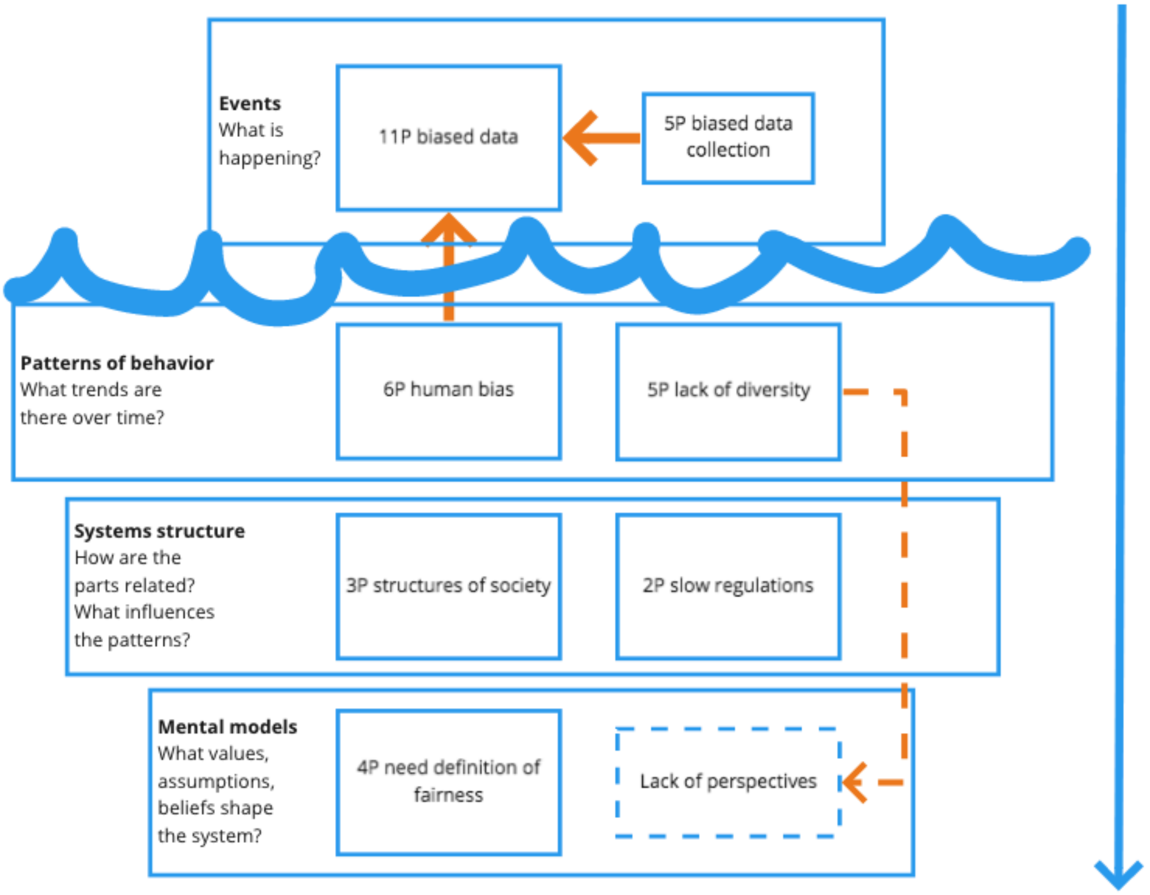


Figure 10. Reported causes sorted into Meadows' Iceberg model. The causes sorted into *Events* are above the water-level and are therefore merely the "tip of the iceberg". Some of the informants point to the cause lack of diversity as a lack of perspectives.

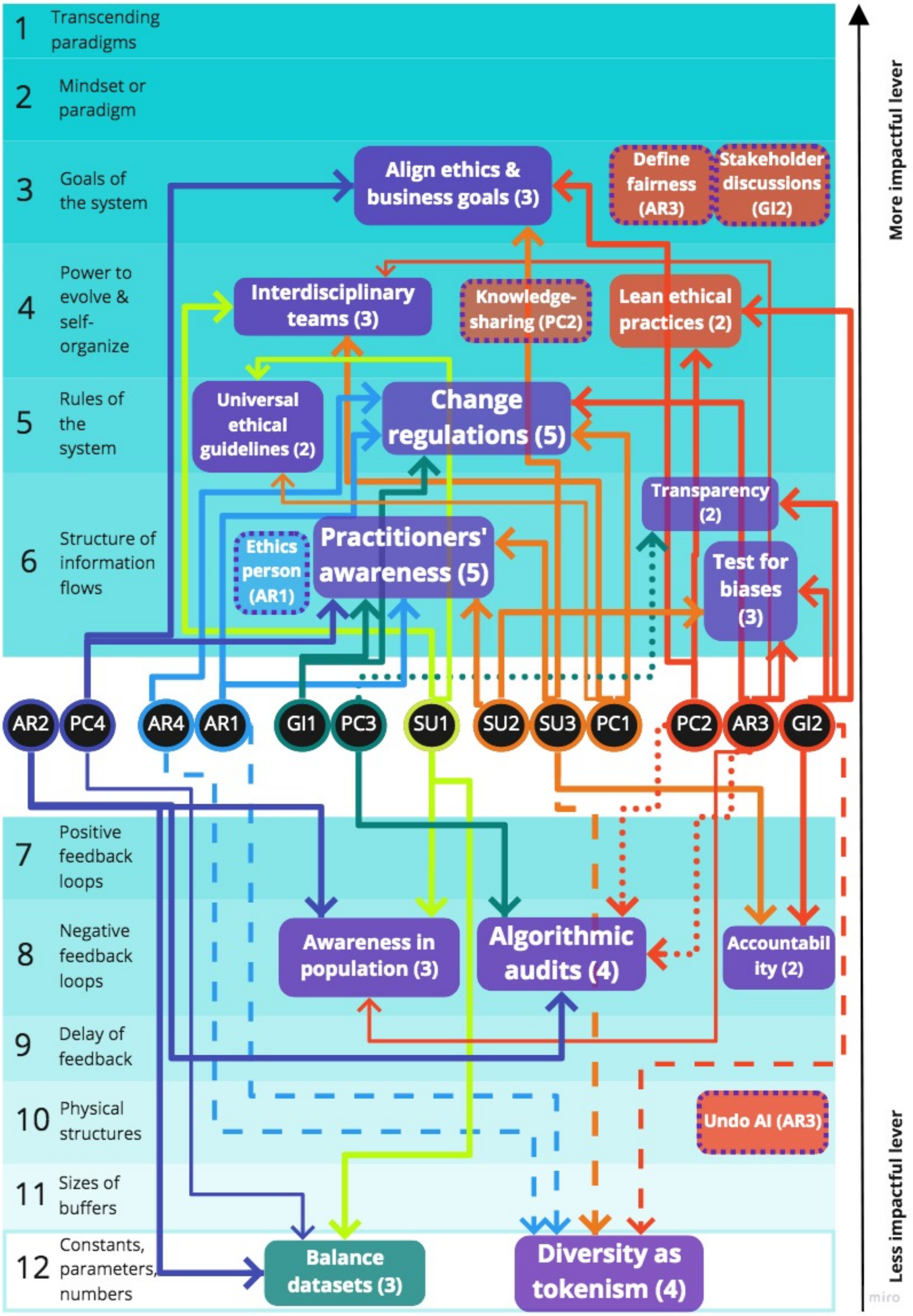


Figure 11. Overview of which solutions was mentioned by which participant, sorted into Meadows' Leverage Points (LP) theory (1999). The names and numbers of the LP are on the left

hand side of the figure. Participants have been grouped together according to similar understandings of technology and ethics practices in groups. The groups are ordered from fewer ethics practices and more Instrumentalist traits on the left, to more ethics practices and some critical perspectives on the right. Those who had some critical perspectives also suggested solutions that were categorized as higher impact.

The boxes represent the different solutions with the number of participants mentioning this solution in parentheses**. Boxes with dotted borders** have only been mentioned by one person and the participant ID is in parentheses, and the color corresponds to the participant**. Red boxes** have only been suggested by red group participants. The **green box** has been suggested by the green and blue groups. Purple boxes have been suggested by combinations of groups from both the left and right hand side.

Although an arrow denotes that a participant mentioned that solution, the participants talked at different depths and lengths about the different solutions. The solutions that they emphasized have thick arrow lines whereas solutions that were briefly mentioned have **thin arrow lines**, e.g. the lines of participant PC4.

Four participants suggested diversity as a solution, this is represented by long **dashed lines** because diversity could be either on a low impact intervention on LP 12 or a high impact intervention on LP 3. That would depend on whether hires of diversity would have the power to also change the goals of the system.

6.2.2 Biased Data: Data is Never Neutral

On a technical and practical level, some of the gender bias is inherited from datasets and standards that are used in the field. One informant talked about how it is required to use the "gold standard" datasets of benchmarking if one wants to be published, despite it being known that the datasets are biased. A researcher's work can't be compared to the work of other researchers if they are not tested against the same dataset, since the performance cannot be compared. However, this relies on the instrumentalist and inherited view that high performance is the valued goal to achieve. According to Bowker and Star, standards can seem neutral and straightforward, but in reality, they are political and "never purely technical" (as cited in Feng & Feenberg, 2008, p. 110). This is confirmed by the informant who states the reason for the gold standard not being changed is not because it is hard to make a new benchmark, but because it requires a lot of resources to convince everybody else to use a new standard; it is a matter of lobbying.

On a deeper level, to point to biased data as a cause and balancing datasets as a solution assumes that there exists an achievable balance where a dataset is neutral. The informants assume that there exists universal knowledge that needs to be found to solve the problem, which is reflected by some of the causes and solutions they suggest: Universal ethical guidelines, balancing datasets, and definitions of fairness.

As mentioned in the previous sub-chapter, the data positivism in the field of AI is a part of the questionable technical heritage from the past. This inherited positivist perspective assumes that there exists objective knowledge that is universally true, that once found it will be true in all circumstances. It presupposes that there is a universally applicable answer that is out there somewhere and just needs to be found. If a definition of fairness that everyone can agree on is found, then all relevant data can be balanced in that manner, the fairness can be programmed and regulated, and guidelines can be made to guide all AI practitioners in the world to achieve that balance.

For instance, if fairness is defined as everyone being treated equally, this would be the equivalent of a facial recognition system working equally well for a man and a woman. This would mean that the dataset would need images that are 50% female and 50% male. And this would be applicable to all image recognition or facial recognition systems in the world. This view of universal balance, fairness, and ethics is too simplistic and problematic on several levels.

First, the reality is more complex than that. A study done by Boulamwini and Gebru (2018) showed that facial recognition systems worked less well for white women compared to white men, but also that the systems worked less well for Black women compared to white men, white women, and Black men. As some of the informants point out, it is mathematically impossible to balance datasets because there are several groups that needs to be balanced, not just the binary genders. If the datasets are balanced separately for race and gender, one risks overlooking the discrimination of colored women (Costanza-Chock, 2018). Without an intersectional view of the problem, one might be creating systems that benefit white women and continue to oppress women of color (Costanza-Chock, 2018; S. M. West et al., 2019). This binary understanding of balance further ignores non-binary individuals on the gender spectrum.

Secondly, when defining fairness it is necessary to ask for whom should the system be fair for, as many people as possible or those whose lives are affected the most? The AI systems don't imitate patterns of sexism and racism because their goal is to find patterns, but because their goal is to find the *dominant* patterns in the data. Finding dominant patterns means that the way decisions have been made for the majority should be applied to everyone else, including the minority. The majority of people have binary genders, so the airport millimeter-wave scanners have binary settings for gender, leading trans people to be flagged due to "anomalies" in the crotch or breast area, and therefore more frequently body searched (Costanza-Chock, 2018). The dominant pattern in Amazon is that men are hired, and therefore the minority of female resumes were filtered aside (Dastin, 2018). These dominant patterns are further exacerbated when data from a more sexist past is used to make systems and decisions for the future, negating the progress that has happened in between (Leavy, 2018). It is doubtful whether it is possible for an AI system that is trained on historical data to be fair for those groups that are erased because they haven't explicitly been counted, like women; or those that there are almost no data on, like women of color or people of LGBTQ+ (Costanza-Chock, 2018; Leavy, 2018; Perez, 2019).

Lastly, what is considered fair is subjective. It is narrowminded to assume that what is considered fair and ethical among genders in countries like Saudi-Arabia or China should also be applicable to Norway, and vice-versa. More importantly, what is considered fair and ethical is not only politically different between individuals with differing political beliefs; a study found that perceived fairness is also different depending on how data and the decision-making processes are presented (Dodge et al., 2019).

Furthermore, as explained in the theory chapter, there is a difference between fairness constituted as equality or equity. Equality gives equal access, but equity takes into account historical inequalities and injustices and aims to compensate for those inequalities so that men and women are equally favored (World Health Organization, 2011). Feenberg (1988) distinguishes these two different notions of unfairness as *substantive bias*, to treat men and women differently when they should be treated equally; or *formal bias*, when women are unfairly disadvantaged because they are treated equally with men. For instance, substantive bias is to filter out female resumes based on gender, whereas formal bias is to filter out resumes based on the most common types of verbs used, which will favor men. Fairness is therefore not so easily concretized to the degree that it is quantifiable, programmable, and always true and the same.

This perspective of universality was also found in the ethnographic studies of AI practitioners conducted in the 1980s and 1990s by Diana Forsythe (2001). In her book *Studying Those Who Study Us*, she points out that the AI practitioners share a positivist understanding of knowledge as something neutral and stable that is somewhere out there and just needs to be found; there is a binary understanding of either knowing and not knowing; and that all the knowing is cognitively located in the mind alone (2001, p. 52). This perspective might lead to a belief that if an answer is found, then the answer is correct; if the system works, then it is correct. But this positivist focus on finding an answer does not question whether there *is* an answer that is always correct.

In contrast, Forsythe (2001, p. 52) describes the understanding social scientists have of knowledge as something that is contextual and hinged on cultural, social, and organizational order, in addition to being a cognitive phenomenon. This view is also supported by Harding who says that all knowledge is partial and limited by its position in history, and limited by the taken-for-granted assumptions of the knowledge (Harding, 1995, 2004). Those with a critical perspective considered factors such as history, culture, context, and society, and therefore suggested more solutions that were higher leverage points.

6.2.3 The Search for the Non-existent Universal Fairness

The instrumentalist understanding of technology enables gender bias to enter AI systems because of its acceptance of error and unexplainable algorithms for the goal of performance and prediction (Jones, 2018). Whereas the positivist understanding of knowledge as universal might contribute to the preservation of bias because the limitations of the different solutions are not sufficiently understood.

Programmed fairness that is algorithmically executed will, first of all, always execute the same model of fairness for everyone. Secondly, as pointed out by one of the informants, the system will "freeze" the way things are done and may prevent future learning for improvement. It is a paradox that AI systems are expected to contain and execute a universal form of fairness when, for example, our legal systems conduct trials to assess the context of each crime. Only one informant mentioned that fairness does not make sense without context. The question AI practitioners need to ask is whether the solution is to find the definition of fairness to code it, or whether alternative ways of executing fairness is needed.

Even the new EU proposal for AI regulation (EU-AIR) does not consider the AI in context with history and society (European Commission, 2021a). In their Questions & Answers section, it is stated that organizations are obligated to ensure that their AIs are properly assessed and audited prior to being put in use, and that this process needs to be repeated when substantial changes are made to the technology (European Commission, 2021b). However, they do not mention the necessity for the AI to be re-assessed and updated as society changes. AI systems also need to evolve in parallel with society's progressive definitions of fairness and those affected by the AI systems, if they are to avoid discrimination.

For instance, facial recognition systems for passport control in Norway need to work equally well for all the ethnicities of Norwegians that come through the airport to avoid discrimination against minorities. Therefore, their accuracy and performance need to be tailored to any additional future ethnicities that previously were not among Norwegians when the algorithm was trained.

Additionally, the passport system’s accuracy needs to change with society, such as the emergence of non-binary identities and expressions among citizens or if new social norms mean previously accepted terms of categorization can become offensive in the future. The new EU proposal for AI regulation assumes either, or both, that AI systems will change often and therefore be audited often; or that once it is "fair" it will *always* be fair.

Even if either of those two conditions were correct, the creators of the proposal seem to be uncertain about what their definition of fairness is. In the proposal, only equality and never equity is mentioned. However, 5 days after publishing the proposal they updated their Q&A where they state: "AI systems can contribute to reduce bias and existing structural discrimination, and thus lead to more *equitable* and non-discriminatory decisions" [emphasis added] (European Commission, 2021b).

Additionally, should the most powerful politicians in the EU be the ones to define what kind of AI fairness is executed in the EU countries? According to Harding, the best view is from the standpoint of the marginalized because they are able to see both their own standpoint in addition to those of the privileged (1995). The EU politicians would be wise to consider Design Justice Principles 2, 3, and 6 as those principles focus on centering the voices of those who are impacted, prioritizes the design's impact on society as opposed to the intentions of the designer, and that "everyone is an expert based on their lived experience" (Design Justice Network, 2018).

6.3 RQ3: What Practices are in Place to Detect and Address Gender Bias in AI?

Few practices are common and in place to address GBAI among the informants. Some of them had more practices, but there were few practices that were regular occurrences. Those who were more affected by the technical heritage had fewer ethics practices than those with a more critical perspective. Few practices are only reasonable if there are no gender biases present in their systems or if they were not developing AI systems, such as one informant who did theoretical research on explainability and did not program AI systems.

However, only one informant's organization actively tested their systems for biases and they were among the few who were aware of what gender bias was present in their data. They were also among those who were developing control points for testing and principles for development. This is perhaps adjacent with the need to translate guidelines into practices that Parsheera (2018) calls for. On the other hand, even if fairness is tested for, whether its results are fair is a subjective opinion that changes depending on how the data is presented (Dodge et al., 2019). And as mentioned, depending on which fairness metrics they use they might be preserving the biases of the status quo (Wachter et al., 2021).

Some of the informants balanced their data or data collection for gender, which are practices supported by the literature (Leavy, 2018; Parsheera, 2018; Zou & Schiebinger, 2018). However, it is not enough for practitioners to balance data on gender in isolation from race if one wants to make AI systems that also benefit colored women (S. M. West et al., 2019). An intersectional approach is required to ensure systems that don't discriminate on gender *and* race (Costanza-Chock, 2018). Some of the informants seemed to have some awareness of the need for an intersectional approach because they mentioned the need to assess which fairness metrics or types of diversity are relevant for a project.

Most of the informants assumed that there were no gender biases present in their systems and did not test for its absence. Some of the informants were not able to imagine how gender bias could be present in their systems; they used word embeddings which were shown to have gender bias but assumed its technical nature would not carry any biases (Bolukbasi et al., 2016); or they implemented AI captioning for videos but did not know that it could work less well for women (Tatman, 2017). These untested assumptions suggest that they might have biases present that they are not aware of.

The lack of bias testing found with the informants coincides with what other research papers claim. Fisman and Luca are cited saying that practitioners don't consider factors such as race and gender and just hope for the best (as cited in Wellner & Rothman, 2020). However, this is problematic because Wellner and Rothman (2020) argue that the odds of AI systems becoming accidentally non-discriminatory are essentially zero. Additionally, if the practitioners are sexist, they might be less likely to perceive biased results as not objective (Otterbacher, Checco, Demartini, & Clough, 2018). This means that it is particularly important to conduct proper bias tests and not just assume that biases are not present because one can't see them.

The informants seem to be preoccupied with whether it works or how well it works, without considering whom it might *not* be working for. True to what an instrumentalist values, they mainly tested for performance and accuracy (Jones, 2018). This is also in line with the five most common values found in AI research: performance, accuracy, understanding, generalization, and building on recent work (Birhane et al., 2020).

Of particular interest was one informant who adamantly asserted that research on differences between genders or ethnicities is "lazy" because one "can always find differences". The informant seems to be missing the point, which is that sometimes there are differences there that should *not* be present and that its presence is an indication of gender bias and inequalities. Additionally, knowledge on gender differences is necessary to mitigate gender biases; if we know that women react differently from men to masculine wording in job ads, we can word ads differently (Gaucher, Friesen, & Kay, 2011). The informant's opinion is in contrast with research which clearly says that gender-disaggregated data and data on protected groups are crucial for the identification and mitigation of inequalities (United Nations University & EQUALS, 2019; Wachter et al., 2021).

Additionally, the research on differences related to minorities and genders is not merely done because it is "interesting", it is sometimes as important as life and death. When there is bias in the data, there are often real life consequences that are felt on the body, like in the example of racial differences of predictive policing or the prediction of recidivism (Richardson et al., 2019). His view however, confirms what another informant said about the lack of diversity: That one of the causes to GBAI is that men are not as interested in solving these problems.

6.3.1 Gender Bias in AI Had Not Even Been Discussed

Several of the informants reported that their organizations had discussed GBAI at varying degrees. It is no wonder that few informants had implemented ethics practices to address GBAI when it had not even been discussed in several of their teams or organizations. Some said that it had been discussed, but they had mainly discussed whether to gender the AI system or chatbot. Although female gendered AI assistants can reinforce harmful stereotypes if they are programmed to be subservient, they are just a small aspect of the issue of GBAI (M. West et al., 2019). The deeper issue is not whether the AI system assistant users interact with are is female, but whether the user is discriminated by the AI because they are female. As mentioned previously, this lack of consideration of societal consequences is a common attitude within the field (Birhane et al., 2020).

6.3.2 Ethics Guidelines

Although many ethics guidelines have been made, of which 84 were analyzed by Jobin et al. (2019), only 4 informants reported that they used a guideline. This appears to be a new finding; studies on whether AI practitioners use ethics guidelines have not been found among any of the 200 references that were added to the reference manager. Those who used guidelines were also more critical and had more reflections and knowledge on the topic.

On the other hand, some of the informants who both did and did not use ethics guidelines, said that it could easily be used for ethics washing; organizations who say they follow certain principles on paper but in reality just do it to look good. It was surprising to find that one informant had made ethics guidelines in one organization but had not implemented any in the organization he worked at, despite being in the leadership team.

This concern related to ethics washing is shared by Wagner (2018) who argues that ethics guidelines are used by the private industry as a means to avoid increased regulation. He says that private companies argue that the ethics guidelines are enough and that they are able to self-regulate because regulation is seen as a barrier for progress (Wagner, 2018). However, one informant argued that ethics guidelines are too vague and that strict regulation is required because AI systems are deterministic systems that can't be self-regulated with hazy instructions.

These opinions are supported by Jobin et al. (2019) and Eitel-Porter (2021). Jobin et al. (2019) found that more information is needed on how to implement ethics guidelines in practice, and Eitel-Porter (2021) says that ethics guidelines are not enough and that proper governance processes overseen by an ethics board are also required for responsible AI practices. This is in line with the practices of one of the informants who was developing new governance frameworks for ethics and implementing an ethics board. Greene, Hoffmann, and Stark are cited criticizing the private industry for using guidelines as a way to make the social problem of ethics in AI seem like a technical issue (as cited in Jobin et al., 2019). According to Jobin et al. (2019), ethics guidelines cannot be implemented using technical skills alone.

Ethics guidelines was suggested by some of the informants as a solution to GBAI. However, as most of the practitioners do not use one and because it is a "soft" type of regulation, its impact might not be that great despite being a LP 5 (see Figure 11). Therefore, a focus on formal regulations might be better, especially since some of the informants expressed that they mainly do the ethics practices that are legally required.

6.3.3 Increasing Diversity

All of the informants but one reported a lack of diversity in their organizations and departments, and three had no female permanent employees. Several of the informants' organizations were thus trying to increase the level of diversity. However, this practice was not done with the purpose of decreasing GBAI, despite several of the informants pointing to diversity as a cause and solution to GBAI. Diversity being the cause of and solution to GBAI is supported by the literature who refers to the current state as a "diversity crisis" (Avila et al., 2018; Leavy, 2018; United Nations University & EQUALS, 2019; S. M. West et al., 2019). Burrell is cited claiming that hiring those who historically have been marginalized will lead to "fair and unbiased" AI systems because prejudices will be decreased (as cited in Timcke, 2020, sec. Encoding Enclosure).

One informant defended their lack of diverse hires with the claim that it is due to the low level of available female talent and having to compete with bigger companies. However, the World Economic Forum reports that the field of AI is not fully utilizing the entirety of the AI talent available and that efforts towards increased inclusion would help solve the issue of unused talent ("Global Gender Gap Report 2020," 2019).

If organizations want to increase their gender diversity, they might have to invest the patience and resources to train women and other marginalized talent on the job. The findings suggest that a change in hiring practices might lead to higher levels of diversity in both gender and nationalities, as the startup(s) who had higher levels of diversity hired employees without AI backgrounds and/or had fully distributed teams. This is similar to how a university in the US managed to increase the number of female students by changing their admissions requirements from programming experience to leadership experience instead (S. M. West et al., 2019).

There might not be any other way around the so-called "pipeline issue" than to hire STEM women that are cognitively capable of learning and doing the job if they receive some additional AI training. If organizations are only interested in hiring candidates that can come do the job without any additional training, diverse talent might end up in other places with those kinds of resources.

Alternatively, the pipeline issue might only increase if women in AI leave the field without women in training to replace them. A report found that nearly half of the women who enter technology eventually leave the field, which is twice as much as the percentage of men who leave the field (Ashcraft, McLain, & Eger, 2016). S. M. West et al. (2019, p. 10) report that not only does AI have a diversity crisis, the trend in both the professional field and computer science education in the USA are going backwards; numbers from 2013 and 2015 are lower than compared to the 1960s and 1980s, respectively.

However, West et al. (2019) criticize the "fix the pipeline" approach. They argue that this approach will not fix AI's diversity crisis because this approach has not created much progress over the last decades. Furthermore, they criticize the pipeline narrative for putting the blame on women and not addressing deeper cultural issues; sexual harassment, wage gaps, power imbalances, and other barriers in "masculine-dominated" institutions that prevent women's success (2019, p. 25). Forsythe supports this claim and says that it is more than a pipeline issue (2001). Both Forsythe (2001, p. XXIV) and West et al. (2019) point to cultural issues and power hierarchies as the reasons for women being forced to "join the silence" (i.e. not be a feminist) or to leave the field of AI, alternatively, not enter it at all. These claims are supported by a study which found that women with degrees in computer science were 14% less likely than men to work in STEM (Sassler, Michelmore, & Smith, as cited in United Nations University & EQUALS, 2019, p. 84).

Surprisingly, in 1983, Forsythe (2001, p. 165) wrote that the field of AI had an "unusually large number of women" compared to the rest of the field of computer science. This is in contrast to the current numbers on women in AI and computer science. Weissman estimates that the whole field of AI only had 13.5% women in 2016, which is lower than the number of women in ICT in the same year; The numbers for women in ICT in Europe (16%) , Americas (22%), and Asia (26%) were all higher in 2016 (as cited in United Nations University & EQUALS, 2019, pp. 85, 96).

The impact of increased diversity would potentially be high or low depending on how it is implemented. Tokenism is defined as when the few employees with marginalized identities are expected to represent the whole identity group; one example of such tokenism is to hire people just to increase diversity statistics (Ashcraft et al., 2016). If diversity is increased in a tokenistic manner without practices for inclusion or allowing women to have a voice, then it is only LP 12, "Constants, parameters, and numbers". Meadows (1999, p. 17) says that changing the hands on the faucet will not lead to any change if the system’s goals are still the same. This is also supported by others who say that the presence of women does not necessarily challenge the systemic structures that lead to the problems in the first place (S. M. West et al., 2019).

An example of increasing diversity without granting the power to change is when Google hired and fired Black female Timnit Gebru (Jonhson, 2020). They muted her research on the inequality of AI language models and then fired her for speaking up about Google's environment of hostility and lack of inclusivity (Jonhson, 2020).

Conversely, increased diversity can be a LP as high as LP3, "goals of the system", if the hired individual is in a position of power and is empowered to change the system goals. For instance, one of the startups had diversity in its leadership and its founders, and they had decided to be an impact company with a double bottom-line consisting of both profit *and* social impact. It is therefore necessary to both include and empower marginalized people *and* to be open to change sole profit goals to include ethics and social sustainability if increasing diversity is to have a significant impact.

However, increasing diversity is about more than increasing perspectives to avoid blind spots. Some of the informants seemed to view diversity of gender and ethnicities as something akin to diversity in perspectives. That if "black 70 year old women" were developing AI systems, then other biases would still be present. According to standpoint theory the view from marginalized black old women would be able to include the views of the privileged white guy (Harding, 2004). So this idea that it's not the white guys that lead to the problems, but a homogenous group making all the decisions, is contested by Harding.

The problem according to standpoint theory, is that when AI is designed from the perspective of the privileged population of white, wealthy, able-bodied, young, straight cis-men, they are not be able to include the knowledge of those who are marginalized (Harding, 2004). According to her, a collection of voices is needed to produce knowledge with strong objectivity (Harding, 1995). The solution is not to "look at things from different angles" but to look at things from the standpoint of those who are most marginalized (Harding, 2004).

Harding's standpoint theory is supported by the finding that practitioners with at least one marginalized identity knew more about the issue of GBAI, and several of them took responsibility to address the issue albeit in varying degrees. Additionally, those with at least one marginalized identity had more critical perspectives (see Figure 12) and suggested solutions that were higher up on LP theory. One informant had no marginalized identities but took it upon themselves to address the issue of biases and the democratization of AI because they had learned more about the ethical issues of AI.

These findings appear to suggest two things. One, that the inclusion of marginalized people might increase awareness, lead to more critical perspectives, and increased ethical practices as a result. Two, that when marginalized perspectives are considered, it is possible for privileged non-marginalized people to understand other perspectives, see more issues, and be motivated to apply more ethics practices in their work. It is worth noting that the non-marginalized informant had learned about these issues from people with at least one marginalized identity. Increasing diversity should therefore be about the empowerment and inclusion of those who are marginalized so that they have a voice to impact the systems that might harm them the most.

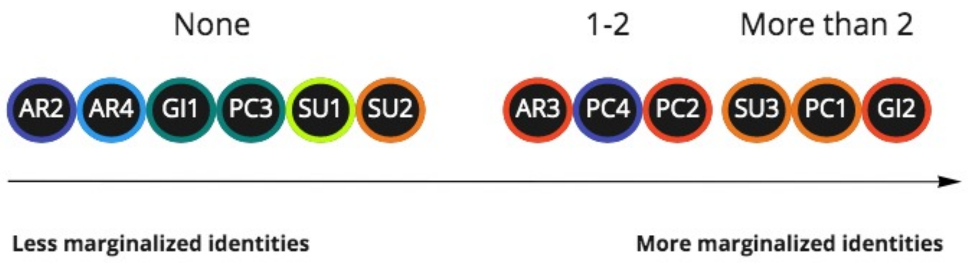


Figure 12. Number of marginalized identities and Critical Perspectives. Informants with at least one marginalized identity had more critical perspectives (orange and red). Informants with no marginalized identities had more instrumentalist perspectives (dark and light blue, dark green) and instrumentalism with some critical perspectives (light green, orange).

6.3.4 Delegating Responsibility

Perhaps the most concerning practice found in the interviews is the delegation of responsibilities. Gender bias is likely to enter AI systems where AI practitioners delegate the responsibility to other institutions and colleagues. A general theme is that the informants don't engage with the whole system, they only focus on what is right in front of them. They forward the responsibility of other parts of the development process to data providers, colleagues, customers, and other entities.

The practitioners delegate the responsibility of gender balance in the data to the data providers or the data collectors when they don't critically assess the data and assume that it is bias-free. They also delegate the responsibility when they continue using biased benchmarking standards without criticizing them in their papers. Even if they *did* balance their datasets, and some of the informants do, it is the lowest leverage point of the suggested solutions (see Figure 11).

The informants seem to have a contradictory belief that practitioners are powerful enough to create systems that are not biased, but at the same time, they are not to blame for biased datasets or biased AI systems. Ruha Benjamin (2019, p. 16) writes that the driving force of oppressive AI systems is that "tech designers encode judgments into technical systems but claim that the racist results of their designs are entirely exterior to the encoding process". She also says that once standards and default settings are created they "take on a life of their own, projecting an allure of objectivity that makes it difficult to hold anyone accountable" (2019, p. 64).

AI practitioners further relinquish their responsibility when, or if, they assume that ethics courses lectured by ethics researchers is enough to prevent future AI practitioners from creating biased systems. Although it is good that they have ethics courses at all, the ethics researchers are probably not teaching the students how to test for biases, which algorithms to use to mitigate biases, or which fairness metrics to use to mitigate the biases of society. Wachter et al. (2021) found that several of the so-called fairness metrics preserve the biases of the status quo.

One informant reported that the organization had a dedicated ethics person, but had not thought of using this person to evaluate other internal projects. Resources had been invested for these ethics people to explore whether AI systems can comprehend morality and ethics, rather than evaluate the ethics of internal projects. It is paradoxical that ethics researchers are paid to find out whether moral choices can be *forwarded* to future AI systems, instead of assessing whether current AI systems are executing questionable moral choices today. If some of the choices that AI systems make today were made by humans, they would have been perceived to be choices of "ethical flavour", according to Dignum (2019a, p. 104).

Dignum (2019b, p. 53) states that there are three principles needed for a "responsible and trustworthy AI": Transparency, accountability, *and responsibility*. It is remarkable that of the three principles, only responsibility is not clearly stated as a cause or solution by the informants. It is concerning that they assume that they are not a part of the problem or contributing to it. Several of the informants had not previously considered whether the AI systems they make might embed gender biases. This is a finding that confirms what Birhane et al. (2020, p. 3) wrote: "We find that overt consideration of societal benefits or harms is extremely rare in our field". The same analysis of Birhane et al. also found that most papers do not reflect on any potential damages and that they mainly focus on performance over ethical principles.

From a business perspective it is always possible for an organization to argue that "we can't afford X", X being any ethics or diversity practices that would require resources. It is also easy to blame their inaction on not knowing how to solve it. Perhaps the companies and organizations who don't have the resources for ethics practices and mitigating biases should not use AI in their systems at all.

However, in the case of the aforementioned informant, the ethics researchers were not utilized simply because they had not thought of doing it. Perhaps this is related to an inherited aversion to consult with other experts instead of just creating things in isolation; this could be similar to the aversion that Forsythe (2001) observed in her studies of AI practitioners. Hopefully this will change. Studies on technology design have shown that actively using ethics people in interdisciplinary teams can function as levers to increase ethics activities, conversations, and procedures (Shilton, 2013, 2014). For instance, in one of the studies, working in interdisciplinary teams lead to more conversations about data, which resulted in conversations about how values like equity need to be incorporated into the design (Shilton, 2013).

Some of them expressed that they only do the ethics practices that are legally required, such as not using non-consensual data from dubious foreign sources or other GDPR regulations; they are in practice forwarding the responsibility to the politicians and the government. The politicians probably know even less about the issues than the AI practitioners, and according to one informant, the politicians are reluctant to touch the subject matter of algorithmic audits.

As an example, the competency of the EU commission is drawn into question because it is unclear whether they understand what they are banning: "AI systems considered a clear threat to the safety, livelihoods and rights of people will be banned. This includes AI systems or applications that manipulate human behaviour to circumvent users' free will (…)" (European Commission, 2021a). Depending on where they have demarcated the borders of responsibility, if they have been defined at all, the new regulation could mean that social media platforms like Facebook will be banned. The ban also depends on how the regulation is interpreted, but some researchers refer to Facebook as a $445 billion dollar company that is among "the most powerful system of manipulating human emotions" that is utilized to turn users into free labor (Jones, 2018, p. 684; Timcke, 2020, p. 9).

Feenberg (2006) emphasizes the importance of democratic interventions in the design of technology, which is also supported by Verbeek (2008) who argues that because the design of technology always has public ramifications the individual alone cannot make such decisions; the design decisions and its ramifications need to be assessed in a public manner. This suggests that people and politicians should be involved in the processes of defining and demarcating what kinds of AI systems we want to allow. However, a study argues that in order for AI to be designed for the greater good, we need to figure out how processes and key stakeholders can be properly informed (Bones, Ford, Hendery, Richards, & Swist, 2020). Voters and politicians need to be able to understand how AI systems work and how AI systems are unable to work (Bones et al., 2020).

Concludingly we can argue that who is responsible for solving issues of gender discrimination in AI and where the problem comes from depends on where one defines the beginning of the AI. Does the problem come from biased data or does it come from society's oppression of women? Is the problem solved when résumé filtering AIs hire women at the same rate as men, or do we also need to change the systems of society? Either way, if AI practitioners believe that they are powerful enough to be able to control whether technology ends up being neutral or not, then they also need to take responsibility for it when it is oppressive. Otherwise, if this is not possible then we need to question whether we should be using them at all.

6.4 Main RQ: What are the Main Perspectives on Gender Bias in AI Among AI practitioners in Norway?

6.4.1 Gender Bias in AI is a Non-Issue For Them

Generally, AI practitioners in Norway do not consider gender bias a prominent issue in their work. Although they acknowledge that the issue is important to solve, most of them assume that their work is not prone to gender biases because their work is different from previously reported cases of biased AI systems, see examples in Background chapter. Several of the practitioners see gender bias in AI as a problem that is not relevant to them because their work is on a technical level that does not involve humans in a way that biases are relevant. Some did not know that it could be an issue in their work. Several informants see gender bias in AIs as an unavoidable issue, but this is worth the benefits of using AIs.

They appear to have not considered their own role and contribution to the issue of gender bias in AI. Some do express an understanding of GBAI and have a somewhat critical perspective on the issue, but few are actively involved in addressing GBAI. As mentioned in chapter 6.1.1, several of the informants have traits of Instrumentalism where they see technology as something neutral and separate from humans and human values. These perspectives and lack of practices could be explained by the effects of technical heritage.

One of the inherited beliefs among the practitioners is the positivist perspective that objectivity is achieved by distancing oneself from one's research and work. Myers (1997) states that positivists assume that "reality is objectively given" and independent from the researcher. Similarly, Harding (1995) argues that impersonal statistical experiments have traditionally been seen as fair and objective. As mentioned previously, this heritage of Instrumentalism and positivism developed from the fields of AI, statistics, and database management after World War II (Jones, 2018).

Harding (1995) criticizes this neutrality ideal of conventional objectivity and states that it is an obstacle to strengthen actual objectivity. Feminist scientists Harding and Haraway assert it is better to see objectivity as a spectrum and divulge its limitations (as cited in Draude et al., 2018), than to operate with the idea of a binary understanding of objective research where the admission of subjectivity is perhaps considered bad research.

The distancing between the practitioners and their work is a sort of deletion of self and social aspects where removing themselves from the equation is seen as achieving objectivity because they think they are no longer affecting their work with social biases. For instance, referring to racial bias in AI systems as "biometrics not well trained", as if there are no social aspects that lead to that bias or any potentially racist humans responsible for that bias. Another example is when they blame the data for biases that might show up in the AIs they make, rather than the reason being *they* did not check the dataset for biases. Similarly, some informants see the reason for not testing for gender biases in their work a result of the *management* not seeing it as an important issue, as opposed to the *informants* not bringing it up to the management as an important issue.

This distancing of self might be a contributing factor to them not seeing their own role and responsibilities on the issue and therefore delegate the responsibilities of addressing it to other entities, or why most of the informants have not considered practices to detect and address GBAI (as mentioned in chapter 6.3.4, Delegating Responsibility). It is therefore important for computerized societies to redefine what responsibility means because the line between cause and effect is less clear, and the fault is often distributed among several actors or even delegated to the technology itself (Nissenbaum, 1996).

Similarly to the distancing of self, another inherited belief among most of the informants is the perception that their AI work and research are apolitical. Their work is perceived as neither addressing gender bias in AI but also not contributing to it. With a few exceptions who *do* strive to address the issue, most informants see themselves as somewhere neutral in between where they focus on making AIs work and optimizing for performance. They also view the AI field and its research as apolitical where biased datasets are a result of unintended unconscious biases or lack of resources, not sexist and racist oppression or veiled discrimination.

This perception that the work of AI practitioners are apolitical was also found by Birhane et al. (2020) and it is also thoroughly rebutted by Abdalla and Abdalla's research (2020). Their study found that about 60% of research on AI and Responsible AI from prestigious universities like Stanford and MIT are funded by Big Tech and that the current tactics of Big Tech funding are similar to how the Big Tobacco industry funded medical research to cover up the negative health effects of tobacco from the 1950s to the 1990s. They state that the tactics employed by Big Tech companies like Amazon, Facebook, Google, Microsoft, Apple, IBM, and OpenAI are "eerily similar" to what Big Tobacco did in the past (2020, p. 9).

Abdalla and Abdalla (2020) found that Big Tech use funding to lobby their interests, improve their public image, and influence research agendas. In addition to funding academic institutions, Big Tech also fund AI conferences like the ACM Fairness, Accountability, and Transparency conference (FAccT) and the Neural Information Processing Systems conference (NeurIPS), both of which have received Big Tech funding every year since 2018 and 2015, respectively (Abdalla & Abdalla, 2020). Their findings refute the belief that AI research or the big AI systems are objective, neutral, or apolitical. Abdalla and Abdalla (2020) suggest increased transparency on funding sources and separation of the field of AI and AI ethics in order to ensure that funding allocation is not skewing the research in favor of Big Tech.

6.4.2 If Only We Were More Aware We Would Make Changes

There is also a pervasive belief among the informants that the main causes to focus on should be the lack of awareness and unconscious human biases. The informants assume that incidents of biased algorithms are unintentional and that all practitioners and organizations need are more awareness to address the issue.

As mentioned in the Theory chapter, Norman said that bad design is made by engineers and managers and that the solution for better design was to create more enlightened designers (as cited in Feng & Feenberg, 2008, p. 106). If the practitioners were powerful, then as Norman believed, the solution would be to enlighten the practitioner about the issues of oppression in AI.

However, the belief that the AI practitioner would do things differently with increased awareness is in contrast to what Feng & Feenberg state (2008). The power analysis of Critical Theory of Technology suggests that knowledge is not enough, it is also about power. Designers do not have full freedom because they are forced to submit to the power relations and hierarchies around them. The effect of such powers are obstacles that inhibit the AI practitioner (Feng & Feenberg, 2008).

This emphasis on awareness is reflected in their suggested solutions because they rely on the assumption that if practitioners knew about the problems then they would avoid them; Increased knowledge is seen as the staple ingredient to make change. These solutions seem to ignore the power relations and goals of profit directing what AI companies and organizations do, and assumes that damaging AI systems are unintentional accidents.

A previously mentioned example of inhibiting power constraints is Google firing first ethics researcher Timnit Gebru for not wanting to rescind her research on the inequalities of large language models, and then firing the AI ethics team lead Margaret Mitchell who criticized Google's decision to fire Gebru (Johnson, 2021). Google confirms Wachter et al. who state that "awareness of inequalities is not the same as rectifying them" (2021, p. 47).

Another example is the standardization of biased performance benchmarks required for publishing AI research, as mentioned in chapter 6.2.2, Biased Data. This discovery that one informant accepted biased standards despite his awareness of the bias was surprising, as there appears to be a dominant belief that practitioners will avoid making biased systems given awareness.

Researchers encourage the use of gender theory to become aware of biases, implicitly assuming that once awareness is reached the bias will be mitigated (Draude et al., 2018; Leavy, 2018). Wellner & Rothman (2020) lists solutions as an evident step after reaching awareness; they appear to not question whether an AI practitioner would need to be motivated to implement a solution. Awareness is probably a necessary step towards a solution, but following CTOT and as suggested by the AI Now Institute, the effects of power relations also need to be taken into account (Feng & Feenberg, 2008; S. M. West et al., 2019).

Increased transparency for the common people as a solution does not take into account the addictive power of algorithmic platforms such as social media. Increased transparency for the goal of market regulation is leverage point 8, "the strength of feedback loops, relative to the impacts they are trying to correct against" (Meadows, 1999, p. 9). Increasing awareness of users through books or speeches as some of the informants do, is not a strong enough lever compared to the ubiquitous presence and immense power of the Big Tech companies.

The informants' assumption that AI companies only need more awareness and that incidents are unintentional is naïve. Abdalla and Abdalla (2020) point out the impersonated innocence of Big Tech in Mark Zuckerberg's apology when Facebook was found to have a role in the tampering of the US presidential election of 2016. In US Congress, Zuckerberg falsely claimed "We didn’t focus enough on preventing abuse and thinking through how people could use these tools to do harm", whereas leaked emails show that they in fact were aware of what Cambridge Analytica did (Abdalla & Abdalla, 2020, p. 2) .

An alternative to awareness as a solution would be regulations because they can level power imbalances in the playing field. Changing regulations is a high-level intervention (LP 5) that could potentially hold companies and organizations accountable. Some of the informants mention regulations and accountability as a solution in addition to knowledge, but generally, most informants do not sufficiently consider the power relations that affects the problem. The informants appear to view the extent of the problem to be limited to practical issues related to the software and its creators. How society affects the design of technology, or how historical patterns of privilege and power determines who are included in a dataset were considered by only some informants with more critical perspectives (see Figure 13).

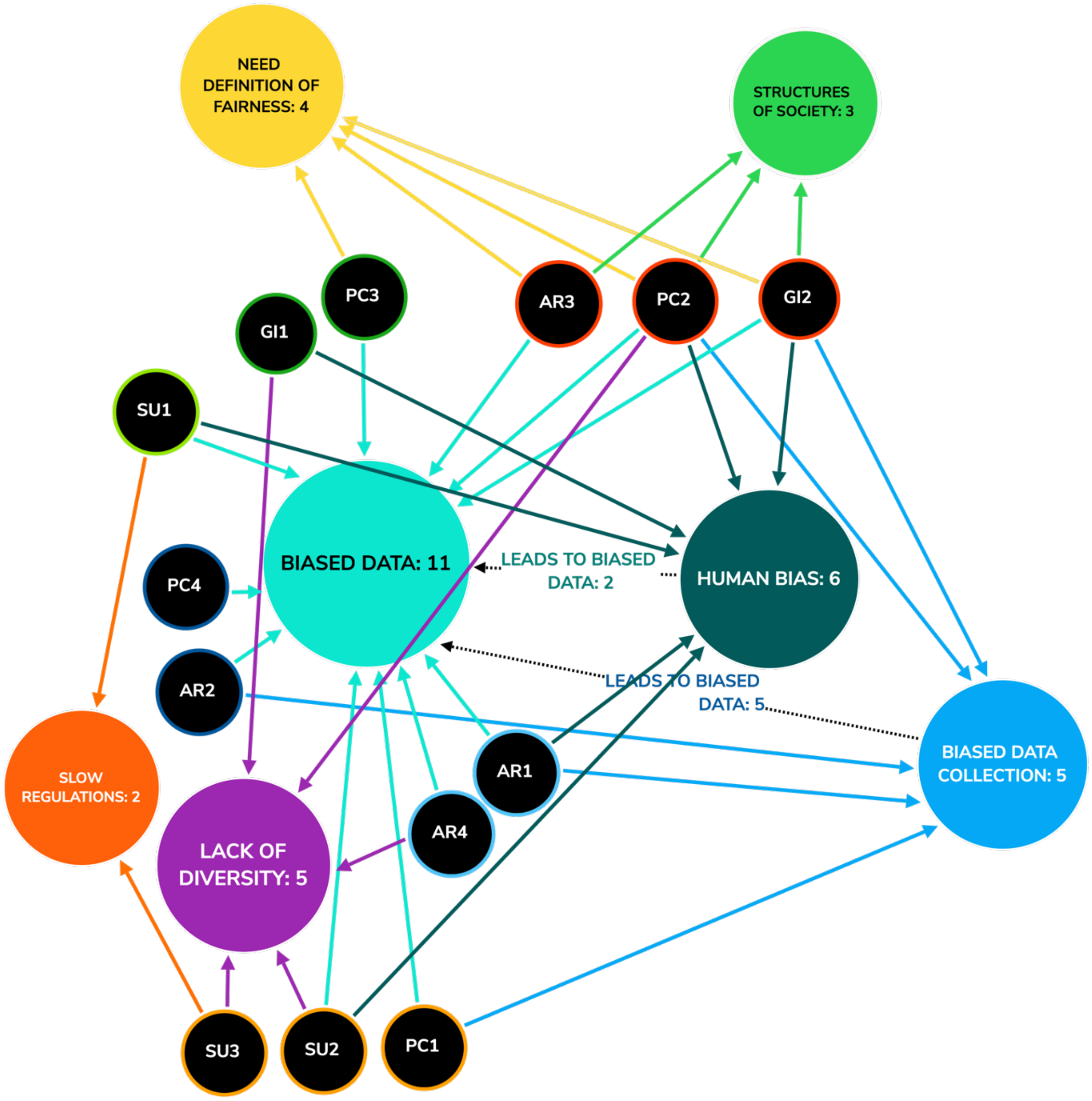


Figure 13. Reported causes of gender bias in AI organized in groups with similar perspectives. The figure shows which causes the different practitioners attributed to gender bias in AI. The different sizes of the circles of causes and the number in the circles denotes how many practitioners said that it was a cause. Full-colored circles are causes, black circles are participants. The colored outlines of the participants denotes the different groupings among the participants. Informants with at least more critical perspectives (orange and red) pointed to more structural causes. Informants with more instrumentalist perspectives (dark and light blue, dark green) and instrumentalism with some critical perspectives (light green, orange) mainly pointed to causes related to data. The black prickled lines show that some causes lead to biased data. Of the 3 who said that human bias was one of the causes, 2 practitioners said that human bias leads to biased data. 5 practitioners said that biased data collection leads to biased data.

However, changing the regulation does not automatically mean that the issue of power imbalance will be addressed if they are not made to protect individuals. The new proposal for AI regulation from the EU Commission does not appear to protect the citizens from the vast powers of governments and Big Tech. The regulations require *users* to flag content that are deep fakes instead of holding content platforms responsible; Law enforcement is exempted from the ban of real-time facial identification systems in public spaces, which might open up for full-time surveillance without the consent of citizens.

Additionally, the EU-AI does not sufficiently empower users so that the power imbalance of Big Tech is adjusted (Timcke, 2020). Completely opting out of AI systems today would mean exclusion from important services and social arenas, therefore citizens and users should have the right to opt out of AI functionalities without receiving negative repercussions from governmental institutions or without having to opt out of using a platform entirely. The regulations also do not require companies to give users ways to challenge outcomes of AI processes. AI companies should be required to implement whistleblower protections that trump other confidentiality agreements to ensure that employees are not fired for exposing unethical AI systems.

More importantly, there appear to be no suggested regulations in the EU-AI to ensure that academic research is not skewed by companies who wish to obscure the damaging impact of AI. AI companies and research institutions should be required to divulge funding relations, and audits should be tasked to assess both algorithms and whether practices aim to be 'ethics washing'.

Although some of the informants see regulations as a healthy limitation and think that some areas should not use AI at all, other informants are more worried about regulation hampering progress. Most of the informants also did not consider removing the algorithm or reverting to analogue alternatives as a solution. The concern that regulation will hamper progress and not suggesting removing AI systems rely on a taken-for-granted assumption that progress in AI is good. This is related to the liberal faith in progress that instrumentalism encompasses.

This perception is also shared by the EU Commission (2021a, p. 10) who wants to "increase people's trust in AI" so as to not hamper innovation and for markets to "flourish". The EU Commission seems to require no assessments of whether AI systems need to produce results of similar quality and equality as the humans they replace. The EU Commission has perhaps been affected by the lobbying of Big Tech as it is the private sector who tends to make ethics guidelines with a focus on "fostering trust in AI" (Jobin et al., 2019, p. 15).

The general approach of many of the research papers found on how to address biases in AI, focus on *how* to fix AI systems and less on *whether* AI systems should be used at all. Timcke (2020) criticizes these approaches as they don't sufficiently consider the systemic forces at play and the embedded ideology of algorithms. According to him, AI systems should not be used where it categorizes people into boxes in a way that "imprison[s]" them with toxic feedback loops. Timcke (2020) uses the example of Facebook marketing high interest loans to users of a lower class who subsequently end up remaining in that lower class because they are not advertised loans with better terms.

It is clear from the examples in the Background chapter that there are many ways in which AI systems can be damaging to those most vulnerable in society. It is not clear that the widespread implementation of algorithmic systems will benefit society, as it appears to mainly bring companies more profit. This is also supported by Timcke, who states that a risk of implementing AI is that its purpose is to "optimize for profit at the expense of people" (2020, sec. Encoding Enclosure).

6.5 Implications for Future Practice and Theory

More practices need to be implemented among AI practitioners in Norway to detect and address issues of GBAI. However, increased awareness, knowledge, or ethics guidelines conflates the bigger issue into a simplistic technical and practical problem; it ignores the problematic power imbalances and lack of responsibility and accountability of Big Tech.

Although it is not irrelevant to attempt to increase intersectional balance in datasets, it should be understood as a quick fix with considerable limitations to what changes in equality or equity it is able to achieve. No dataset is ever neutral because a universal fairness does not exist. Attempts at decreasing discrimination in AIs should therefore aim to protect those who are most vulnerable and disadvantaged in society.

Increasing diversity should not be seen as a solution to fix the pipeline or to increase the number of perspectives, rather it should be seen as the empowerment and inclusion of those who are marginalized so they have a voice to impact the systems that might harm them the most. Those who have no marginalized identities need to take the time to learn about the experiences of those who are marginalized in order to better understand marginalized perspectives and strengthen their level of objectivity.

The AI field has evolved from simple ones and zeroes and making AIs that can beat chess champions (De Spiegeleire et al., 2017) to developing human systems: systems that sort human lives. Human lives are in the hands of AI practitioners but end user agreements ensure that nobody is responsible or liable for any damage to those lives (Nissenbaum, 1996). Human systems is a territory that is not sufficiently understood without the consideration of power relations because as Kimberle Crenshaw states, "the process of categorization is itself an exercise of power" (as cited in Timcke, 2020, sec. Data Politics and Knowledge).

The field of AI has emerged from positivist fields of applied statistics and data management (Jones, 2018) but has developed into something closer to the social sciences or medical technology. More research is needed to map the technical heritage of AI, on how it contributes to GBAI, and which alternative future paths are possible. AI practitioners need to be aware of the inherited assumptions that they take for granted and whether these are assumptions that are beneficial to keep.

Feenberg's (2006; Feng & Feenberg, 2008) philosophies of technology and instrumentalization theory, and also Harding's (1992, 1995, 2004) theory on standpoints and strong objectivity have been useful for understanding the views of the informants. A re-evaluation of the philosophy the field decides how knowledge is known and how objectivity is maximized is certainly needed. Objectivity should be seen as a spectrum that is accounted for, rather than a binary presence or absence because the deletion of the self appears to lead to a delegation of responsibility. At the very least, students who become future AI practitioners need to be presented with alternative philosophical approaches in order to make informed decisions on how to view the world and understand its limitations.

Meadows' LP theory (1999) has been useful for comparing causes and solutions and theories on how change is made should be used in AI to determine which solutions to implement. One of the highest levers of intervention (LP 2) would be a paradigm shift from instrumentalism and positivism, to critical perspectives encompassing CToT, feminist standpoint theory, intersectionality, and Design Justice Principles. A more critical perspective would likely be better adept at understanding the full extent of the issue and therefore also be able to see interventions that are of higher impact. It is possible to have a critical perspective and still optimize for technical accuracy, but it is not likely that the systemic, political, historical, and social nature of gender bias in AI will be solved through "neutral" technical solutions. Non-neutral, perhaps biased, solutions would be to listen to marginalized views as the Design Justice Principles suggest (Design Justice Network, 2018), learn from social and feminist scientists, and politically position one's work to intentionally benefit those less privileged instead of enabling existing systems of oppression.

7 Reflection

7.1 Interacting with the Views and Power of the AI Practitioners

I found it surprisingly that gender alone did not explain the level of awareness on the topic of GBAI; the size of an organization or level of resources did not explain the level of ethics practices. Finding out that there is a lack of ethics practices was expected, yet, the findings were still disappointing. It was especially disappointing and shocking when the informants' comments clearly contrasted my prior assumptions about the issue, such as finding that practitioners can be aware of biases and still not address them. It was also surprising to learn that it is harder to solve the issue than originally expected because universal forms of fairness are mathematically impossible, and the issue is much larger than what the individual practitioner chooses to do.

It was challenging to interview on such a sensitive topic as gender and racial bias with male and white informants and write a critical perspective on their lack of ethics practices. The informants are my potential future colleagues, peers, and maybe even employers, therefore our interactions during the interview could affect my future career in AI in Norway. Informants were sent a copy to review prior to submission which gave them an opportunity to ask for changes to be made. However, this was also nerve-wracking as my interpretations could lead to misunderstandings which they could potentially find offensive.

Ways to avoid this dilemma could for instance be that someone outside the field interviewed them, such as an ethnographer. However, then this person might lack the level of technical understanding needed to probe into some of the topics. Another alternative could have been to ask informants to clarify findings that could be interpreted in alternative ways or to have another researcher code the transcribed interviews to compare interpretations. This is something Braun and Clarke recommend that one does (Braun et al., 2018).

7.2 Different Use of Theory

I had just finished my interviews, when I came across a newly published book called Data Feminism. Data Feminism is an approach to data science coined by researchers Lauren Klein and Catherine D'Ignazio (D’Ignazio & Klein, 2020). Their recently published book encompasses several of the traits of theories used, such as standpoint theory, intersectional feminism, Design Justice, and Critical Theory of Technology. Although the theories chosen for this thesis were well-suited for the analysis and discussion, Data Feminism could have provided a more coherent framework with additional aspects not considered here. For instance, it could have been interesting to investigate the informants' perspectives on concepts that the authors call Imagined Objectivity versus concepts such as justice.

With a perspective that emphasizes the role of power, it would have been more interesting to explore what organizations with power do to incentivize others to address the issue. Such organizations could have been funding institutions like The Research Council of Norway or Innovation Norway.

7.3 Limitations

The findings on perspectives are limited by my interpretation of what the informants said about other topics because they were not explicitly asked about their assumptions or beliefs. This is because the study had an exploratory approach and such findings were unexpected, they were therefore not explicitly asked. Their perspectives could have been validated through a longer follow-up interview or a follow-up survey, but this would have required more resources. However, asking about their understandings on technology in terms of instrumentalism might have been difficult because it would have required a common understanding of what it means to view technology as neutral.

Triangulation of Data

Some triangulation of data collection was done; diversity statistics of the informants' organization were confirmed by checking their organizational annual reports (Yin, 2009). Some theory triangulation was done; the interpretation of some of the data was discussed with the thesis supervisor to get a second opinion (Yin, 2009).

Validity and Reliability

Validity was increased by using theories (Yin, 2009): Feenberg's philosophies of technology and Harding's standpoint theory were used to understand the perspectives of the informants; Meadows' leverage points theory was used to assess the impact of causes and levers of intervention.

The reliability of the study was increased by creating a case study database in Nvivo and maintaining a chain of evidence from data collection, then transcription, and importing all relevant documents in the database (Yin, 2009).

Respondent Bias

Answers from informants are likely to be affected by the behavior of the researcher and who the researcher is (Myers, 1997; Myers & Newman, 2007). Particularly, answers to questions about gender bias conversations on racial bias, might have been skewed because I am a young woman with Asian heritage (M. D. Myers & Newman, 2007). Additionally, their answers might have been distorted because they wanted to look good or please the interviewer (Robson, 2002). Cues to what answers were desirable were minimized by mainly asking open-ended questions, especially in the beginning, and by introducing the project topic in a high-level manner (Robson, 2002; Taylor et al., 2016), i.e. the participants were asked to participate in a research project on "AI ethics" rather than "gender bias in AI". To encourage informants to speak more openly, advice from Robson (2002) was followed; listen more than speaking; ask questions in a non-threatening manner; give the impression that there was no judgment and that it was an enjoyable conversation.

Additionally, the interview skills improved over time which could have introduced a difference between the early and later interviews in the resulting answers. This bias was decreased by conducting two test interviews. Although measures were made to minimize how informants were affected by researcher behavior, there are no guarantees that they were completely eliminated.

Researcher Bias

Other researchers should consider using techniques from social science which have a tradition of questioning which biases they bring into a study (Crawford, 2013). Researcher identity memos where assumptions, beliefs, and expectations are explored prior to a research project could be a useful mitigation technique (Maxwell, 2012).

7.4 Imagined Objectivity and Data Feminism

Both the literature and the informants talked about balancing datasets as if neutrality or objectivity are opposites of bias. I initially shared this view until some informants pointed out the mathematical impossibility of fairness and I came across Data Feminism. Data Feminism succinctly describe the misguided practices of searching for a point of objectivity as a solution to bias as *imagined objectivity.*

Klein and D'Ignazio (2020) elaborate on the concept of imagined objectivity which was originally coined by sociologist Ruha Benjamin (2019). D'Ignazio and Klein (2020) argue that it is necessary to approach issues of GBAI with a point of view that objectivity in systems are not real. They explain that cultural assumptions lead to ideas of imagined objectivity in systems because they are seen as technical and therefore assumed to be less partial and discriminatory. They say that "all systems are political", and political means they are not objective (D'Ignazio & Klein, 2020, p. 62). The authors say that it is not enough to look at the systems. We also need to consider the culture, history, and context that shaped those systems in the first place. Even though I was spurred on by S. M. West et al. (2019) to look at how power affects these issues, Data Feminism argues it is not enough to only look at the issues of power and bias within the bounds of the AI field without considering imagined objectivity (D'Ignazio & Klein, 2020).

In the debates and ethics guidelines of biases in AI, and also in my study, certain words keep reappearing: Ethics, bias, fairness, and understanding algorithms, also called explainability (Jobin et al., 2019). Klein and D'Ignazio criticize the use of such words because they encompass imagined objectivity; the concepts point to no one in particular in terms of who is responsible. At most, concepts of imagined objectivity put the blame on individuals or technical systems for bias issues. Alternative concepts which challenge systemic power differences are justice, oppression, equity, and understanding history, culture, and context, see Table 17.

The left column of table 14 includes terms and concepts that many ethics guidelines include (Jobin et al., 2019) and who dominate the conversations around the issues of discriminating AI systems (D'Ignazio & Klein, 2020). **Bias** refers to bias in people, algorithms, or datasets, but does not account for history, context, or culture. **Restorative justice** accounts for the historical damages. **Fairness** lacks context and history, does not acknowledge the systemic nature of discrimination. **Equity** accounts for history and differences in power of different groups. **Understanding algorithms**; D'Ignazio & Klein refer to this as a good start but understanding how machine learning works is not enough. Understanding **history, culture, and context**; we need to understand how discriminating systems are shaped by history, culture, and context. Looking at history focuses solutions on how we end up with an oppressive AI system instead of only how to make them "more fair".

Table 17
Concepts of Imagined Objectivity from Data Feminism

Concepts Which Uphold "Imagined Objectivity" and Secure Power Because they locate the source of the problem in individuals or technical systems	**Intersectional Feminist Concepts Which Strengthen Real Objectivity and Challenge Power** Because they acknowledge structural power differences and work towards dismantling them
Ethics	**Restorative justice**
Bias	**Oppression**
Fairness	**Equity**
Understanding algorithms	**Understanding history, culture, and context**
Note. Adapted from Data Feminism, by C. D'Ignazio and L. Klein, p. 60. Copyright 2018, 2020 by MIT Open Press. Reproduced with permission.	

D'Ignazio & Klein argue that concepts in the left side of the table are not enough to address the deeper issues of inequality without the concepts on the right. They argue that the left hand concepts might be useful to solve technical systems in isolation, but they are not enough to address the problem at the root cause of systemic injustices. A non-discriminating resume filtering AI does not solve Amazon's systematic pattern of hiring less women, and it does not solve the cultural problems in computer science causing women to leave the field (Forsythe, 2001; United Nations University & EQUALS, 2019; S. M. West et al., 2019). "Bias" puts the blame on the AI system and Amazon was therefore let off the hook when they scrapped the biased resume filtering AI. Solutions of imagined objectivity seem alarming because AI companies can continually make biased systems and might only need to scrap the systems every time to claim they have "solved" the problem.

As discussed in chapter 6.2, there exists no universal form of ethics or fairness because it is a mathematical impossibility, and what is considered ethical or fair is contextual, subjective, and political. Therefore, it is better to aim for restorative justice and equity for those who historically have been marginalized. When it is not possible to make systems equally fair for all genders and all races, then systems of justice can be made to tip the scales to favor those whose odds are usually against them. A system where fairness is equality would hire based on technical skills and provide bias training for men; equity is valuing the lived experiences of women as an asset and balancing structural barriers in education and culture by providing women training in AI.

Klein and D'Ignazio state, biases are not accidental phenomena that mysteriously end up embedded in AI systems; they are symptoms of oppressive systems that benefit and favor certain privileged people, and they are diversions from systemic issues. AI systems are not biased against women, they oppress women; from entering jobs, through oppressive resume filtering; utilizing technology, through oppressive facial recognition; and breaking stereotypes of what women are able to do, with oppressive representations of women; or how they should behave, by encoding oppressive voice assistants who reinforce the idea of the docile woman.

For instance, the cause of GBAI is not explained by bias in benchmarking datasets alone. The TIMIT benchmarking dataset, where only 30% of the recordings are of women (Rémy, 2021), is likely a reflection of the gender parity at MIT during its conception in the 1990s; 35% and 19% female bachelors' and masters' graduates in 1995, respectively (MIT Institutional Research, n.d.). This again may be a symptom of the privileges of who got to enroll and the oppression of those who did not, harking back to systemic factors like gender, race, class, and access to resources.

The imagined objectivity is visible in the informants' perception that their work is apolitical. They see biases as accidental problems that can be mitigated with information, and they think their inaction on the issue is a neutral, not enabling, stance. Those who considered the aspects on the right side of the table had a deeper understanding of the problem. Some of the informants work to increase the explainability of algorithms. Although it is important to understand how algorithms work and to make them explainable, it is not enough for providing justice. Understanding history, culture, and context are necessary to assess what justice means and for taking into account historical injustices. Without the cultural knowledge on how women use different verbs in resumes, it is not possible to prevent female resumes from being filtered out.

Klein and D'Ignazio (2020) call for us to question, *who* does the system benefit and *who* does it harm? Who does concept of imagined objectivity benefit and who does it harm? It benefits the companies who do not want to be held liable as an institution. It benefits organizations who claim 'we fixed it' when they have an AI system that can check the fairness of other AI systems or when they can point to hard numbers and say 'our systems are fair because everyone is treated equally' without considering that everyone does not start out with equal privilege and power. Practical solutions solve superficial causes but they omit root causes of oppression, which leads to problems being repeated in other ways (D'Ignazio & Klein, 2020). Sometimes I wonder why AI systems are used when they appear to benefit those already in power more than everybody else.

Words have defining powers over how we think and understand problems. Bones et. al (2020) state "We raise a concern that some of the ethical problems now presented by AI (and highlighted by many scholars) are a product of the language used." It is therefore important to consider the social implications of which words are chosen to describe the issues of female oppression in AI. The words define how the problems are addressed and when the goal is reached. In the Lexin English-Norwegian dictionary equity is translated into "lik", equal; and fairness and justice converge into the same word, "rettferdighet" (LEXIN, 2019). The fact that we do not even have Norwegian words which differentiate fairness from justice and equality from equity says something about the current state of debate and progress. More importantly, the lack of term hampers discussions and awareness because progress would require lectures on terminology.

There is nothing wrong with being a white guy in AI. There could be something problematic with having a privileged point of view and pretending like it is an objective point of view from where research should originate. During this project, I have realized more and more that research is not objective and neither is this thesis project. My subjectivity – stemming from my negative experience with algorithms that treat everyone the same – is a resource, a window into the world of the oppressed, it is not an unwanted bias of this thesis. The point is that because research is not 100% objective it cannot be viewed or claim to be as such. There can be degrees of objectivity and research might strive for as much objectivity as possible, but it does more harm than good to falsely present research as a beacon of neutrality and objectivity when in reality it is affected by history, culture, context, and systemic injustices by those in power.

The issue with oppressive AI systems is like the discussions around abortion; Abortion is murder depending on where one defines the beginning of a life. With AI, who is responsible for solving issues of gender discrimination in AI and where the problem comes from depends on where one define the beginning of the AI. Does the problem come from biased data or does it come from society's oppression of women? Is the problem solved when translations include she as often as he or do we also need to change something else?

Publicly funded researchers and government employees have a public duty to make sure that citizens are not harmed or oppressed by the systems they pay for. With great power comes great responsibility. AI is perhaps the greatest power mankind has ever had but no one who is wielding it seems to be taking the responsibility for its damaging impact.

7.5 Future Work

Future work could compare the effects on ethics of different suggested solutions and levers of intervention. For instance, how a paradigm shift to critical perspectives would affect and benefit AI projects; how an AI project team with at least 35% marginalized individuals might compare to a less diverse AI project team; how the power of Big Tech can be more evenly distributed through measures such as whistleblower protections and worker's unions; or the potential effects of a future implementation of EU-AIR.

Future work should aim to do a study to explore what other technical heritage there is, to which extent the heritage is present in the AI field in Norway, and how that affects the oppression of marginalized groups in AI projects.

Looking more broadly, future work could examine enabling power structures such as funding sources. Aspects that should be examined are their ethical focus areas, whether they try to mitigate bias issues in AI, and what kinds of funding policies would be beneficial to mitigate bias issues in AI.

Women have a greater risk at losing their jobs due to AI automation compared to men and as many as 11% of the jobs held by women today may be gone in the future (Mohla, Bagh, & Guha, 2021). Future research on GBAI should therefore examine how the AI industry affects the labor forces that are currently dominated by women.

8 Conclusion

This qualitative multiple case study research has interviewed thirteen practitioners from the AI field in Norway. It explored their perspectives on gender bias in AI, what understandings of technology they have; how gender bias enters an AI system; and what practices are in place to detect and address gender bias in AI.

The technical heritage of computer science and AI has an important influence on AI practitioners' understanding of technology. The informants in this study showed traits of instrumentalism and separated themselves from the technology they create, because they think the separation leads to objectivity. Some informants showed a more critical perspective on how research and AI ethics practices are done or have been done, and they considered technology in relation to themselves, society, history, power, and other people. Power is an important factor that practitioners need to consider to solve gender bias in AI at a systemic level.

AI practitioners in Norway seem to have insufficient knowledge on how gender bias enters an AI system. A contribution of this thesis is a figure that depicts the entry points of gender bias in AI based on the literature review. Compared to the reported causes, the informants did not recognize the same number of entry points as found in the literature. Most of the informants could only identify a few of the entry points, and the most mentioned entry points were somewhat superficial entry points like biased data, human bias, and lack of diverse perspectives. The practitioners' belief that there exists a form of fairness that will always be correct for everyone at all times without considering the context, enables biases to enter AI systems.

More critical perspectives among AI practitioners are needed to understand and address gender bias in AI. The less critical informants reported more superficial causes and solutions of lower impact to address gender bias in AI. Using Meadows' levers of intervention, causes and solutions were categorized and assessed for level of impact; a comparison of the impact of solutions like this has not been seen in previous literature. A commonly suggested solution among the practitioners and literature is the low-impact lever of intervention of balancing datasets. Biased data are a superficial low-impact lever of intervention, and the notion of a neutral dataset does not exist because universal fairness is mathematically impossible. Practitioners should instead focus on addressing systemic issues of power and aim for AI systems that incorporate justice for those who historically have been marginalized.

The positivist heritage of the field limits the perspectives of AI practitioners in Norway because they have a technical view on the problem and underestimate the role of power. One of the limiting positivist beliefs is that there exists universal knowledge that when found would solve the problem: universal ethical guidelines, balancing datasets, and definitions of fairness. However, this type of universal knowledge does not exist because definitions of fairness and ethics are contextual, subjective, and political. They have an instrumentalist assumption that progress in AI is good and a naïve perception that the main issue is a lack of awareness and unconscious biases. Most of the solutions suggested by informants assume that if practitioners knew about the problems, they would avoid them, but some of the informants were aware of biases and did not address them. Furthermore, literature indicates that big tech companies are more than aware of ethical issues in AI, and they lobby to prevent regulation and deeper awareness.

Several of the practitioners saw gender bias in AI as a problem that was not relevant to them or a great risk in their work. They appear to have not considered their own role and contribution to the issue of gender bias in AI. One of the taken-for-granted assumptions is the positivist perspective that objectivity is achieved by distancing oneself from one's research and work. This belief may be the reason that practitioners distance themselves from the problem of gender bias in AI and therefore delegate the responsibilities of addressing it to other entities and institutions. The field of AI needs to re-evaluate its research philosophy and examine what technical heritage and taken-for-granted assumptions are negatively impacting the research on gender bias in AI.

The delegation of responsibility meant the practitioners implemented a limited number of practices: Use of ethics guidelines, testing for biases, balancing for gender in datasets or data collection, and increasing diversity. More practices need to be implemented to detect whether gender biases are present in AI systems and to address the problem.

The informants with at least one marginalized identity knew more about the issue, and several of them took responsibility to address the issue. Those with no marginalized identities should strive to understand and learn more about marginalized experiences to strengthen their objectivity. Therefore, increasing diversity is an essential solution. However, the solution is not about increasing the number of perspectives but including and empowering marginalized voices. Only from the standpoint of those marginalized can objectivity be strengthened because marginalized perspectives also understand privileged views. Therefore, hiring practices should change and increase diversity by training disadvantaged groups in AI, rather than giving bias training to non-marginalized people.

This study shows that a paradigm shift is needed from instrumentalism, positivism, and imagined objectivity to a critical, intersectional perspective that includes and empowers marginalized voices. Practitioners need to understand that their work *is* political, and doing nothing to address the issue of gender bias in AI is enabling the harm that is done. Common terms in the conversations on AI like ethics, fairness, bias, and explainability do not address the systemic root causes. Practitioners should instead focus on words like justice, equity, and oppression in order to challenge power structures and fix the problem at its root.

References

80,000 Hours. (n.d.-a). About us. Retrieved April 28, 2021, from 80,000 Hours website: https://80000hours.org/about/

80,000 Hours. (n.d.-b). Our current list of the most important world problems. Retrieved January 21, 2021, from 80,000 Hours website: https://80000hours.org/problem-profiles/

Abdalla, M., & Abdalla, M. (2020). The Grey Hoodie Project: Big Tobacco, Big Tech, and the Threat on Academic Integrity. *Resistance AI Workshop @NeurIPS 2020*. Presented at the Thirty-fourth Conference on Neural Information Processing Systems, Virtual. Retrieved from https://sites.google.com/view/resistance-ai-neurips-20/accepted-papers-and-media

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–18. Montreal QC Canada: ACM. https://doi.org/10.1145/3173574.3174156

Adams, Rachel. (2020, March). Artificial Intelligence has a gender bias problem—Just ask Siri. Retrieved January 22, 2021, from Human Sciences Research Council website: http://www.hsrc.ac.za/en/review/hsrc-review-march-2020/artificial-intelligence

Adams, Richard, & McIntyre, N. (2020, August 13). England A-level downgrades hit pupils from disadvantaged areas hardest. *The Guardian*. Retrieved from https://www.theguardian.com/education/2020/aug/13/england-a-level-downgrades-hit-pupils-from-disadvantaged-areas-hardest

Algorithmic Justice League. (2020). Algorithmic Justice League. Retrieved January 27, 2021, from https://www.ajlunited.org/

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Arnold, M., Piorkowski, D., Reimer, D., Richards, J., Tsay, J., Varshney, K. R., … Olteanu, A. (2019). FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*, *63*(4/5), 6:1-6:13. Scopus. https://doi.org/10.1147/JRD.2019.2942288

Ashcraft, C., McLain, B., & Eger, E. (2016). *Women in Tech: The Facts* (p. 76). National Center for Women & Information Technology.

Avila, R., Brandusescu, A., Freuler, J. O., & Thakur, D. (2018). *Policy Brief W20 Argentina—Artificial Intelligence: Open questions about gender inclusion*. Argentina: World Wide Web Foundation. Retrieved from World Wide Web Foundation website: http://webfoundation.org/docs/2018/06/AI-Gender.pdf

Bates, J., Clough, P. D., Jäschke, R., & Otterbacher, J. (Eds.). (2018). *Proceedings of the International Workshop on Bias in Information, Algorithms, and Systems (BIAS)*. Aachen. Retrieved from http://ceur-ws.org/Vol-2103/

BBC. (2020, August 20). A-levels and GCSEs: How did the exam algorithm work? *BBC*. Retrieved from https://www.bbc.com/news/explainers-53807730

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., … Zhang, Y. (2018). *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*. Retrieved from http://arxiv.org/abs/1810.01943

Benjamin, R. (2019). *Race after Technology: Abolitionist Tools for the New Jim Code*. Medford, MA: Polity.

Bergsjø, L. O., & Bergsjø, H. (2019). *Digital etikk: Big data, algoritmer og kunstig intelligens*. Oslo: Universitetsforlaget. Retrieved from https://www.universitetsforlaget.no/digital-etikk-1

Birhane, A., Kalluri, P., & Card, D. (2020). The Underlying Values of Machine Learning Research. *Resistance AI Workshop @NeurIPS 2020*, 6. Virtual: Resistance AI Workshop @NeurIPS 2020. Retrieved from https://sites.google.com/view/resistance-ai-neurips-20/accepted-papers-and-media

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 29). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

Bones, H., Ford, S., Hendery, R., Richards, K., & Swist, T. (2020). In the Frame: The Language of AI. *Philosophy & Technology*. https://doi.org/10.1007/s13347-020-00422-7

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., Clarke, V., Hayfield, N., & Terry, G. (2018). Thematic Analysis. In P. Liamputtong (Ed.), *Handbook of Research Methods in Health Social Sciences* (pp. 1–18). Singapore: Springer. https://doi.org/10.1007/978-981-10-2779-6_103-1

Brombach, H. (2016, March 8). Kvinneandelen i den norske IT-bransjen er svært stabil. Retrieved January 26, 2021, from Digi.no website: https://www.digi.no/artikler/kvinneandelen-i-den-norske-it-bransjen-er-svaert-stabil/348237

Brownell, P. (2010). Chapter 1—Social issues and social policy response to abuse and neglect of older adults. In G. Gutman & C. Spencer (Eds.), *Aging, Ageism and Abuse* (pp. 1–15). London: Elsevier. https://doi.org/10.1016/B978-0-12-381508-8.00001-1

Bubakr, H., & Baber, C. (2020). Using the Toulmin Model of Argumentation to Explore the Differences in Human and Automated Hiring Decisions. *Proceedings of the 15th International Joint*

Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 2: HUCAPP, 211–216. Science and Technology Publications, Lda (SciTePress). https://doi.org/10.5220/0009129102110216

Buolamwini, J. (2016). *How I'm fighting bias in algorithms* [Video file]. TED. Retrieved from https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms

Buolamwini, J. (2018). *AI, Ain't I A Woman? - Joy Buolamwini*. youtube.com. Retrieved from https://www.youtube.com/watch?v=QxuyfWoVV98

Buolamwini, J., & Gebru, T. (2018). *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification* (A. F. Sorelle & W. Christo, Eds.). Proceedings of Machine Learning Research: PMLR. Retrieved from http://proceedings.mlr.press/v81/buolamwini18a.html

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Celis, L. E., Huang, L., Keswani, V., & Vishnoi, N. K. (2019). Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 319–328. Atlanta GA USA: ACM. Scopus. https://doi.org/10.1145/3287560.3287586

Coded Bias—Global Release Marketing Toolkit. (n.d.). Coded Bias. Retrieved from https://www.codedbias.com/s/CODED_NETFLIX_Toolkit_Final.pdf

Costanza-Chock, S. (2018). Design Justice, A.I., and Escape from the Matrix of Domination. *Journal of Design and Science*. https://doi.org/10.21428/96c8d426

Crawford, K. (2013). *The Hidden Biases in Big Data*. Retrieved from https://hbr.org/2013/04/the-hidden-biases-in-big-data

Crawford, K. (2016). *Artificial Intelligence's White Guy Problem*. Retrieved from https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html

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

Data2x. (2020). *Talking Data Feminism with Catherine D'Ignazio and Lauren F. Klein* [Video file]. Data2x. Retrieved from https://www.youtube.com/watch?v=pmNEe6FvduM&t=6s

Datatilsynet. (2020, May 26). Starter regulatorisk sandkasse for utvikling av ansvarlig kunstig intelligens. Retrieved October 23, 2020, from https://www.datatilsynet.no/aktuelt/aktuelle-nyheter-2020/regulatorisk-sandkasse-for-utvikling-av-ansvarlig-kunstig-intelligens/

De Spiegeleire, S., Maas, M., & Sweijs, T. (2017). *WHAT IS ARTIFICIAL INTELLIGENCE?* (pp. 25–42). Hague Centre for Strategic Studies. JSTOR. Retrieved from Hague Centre for Strategic Studies website: www.jstor.org/stable/resrep12564.7

Dejan Jotanovic. (2018, June 19). This is how artificial intelligence is undoing women's rights. *The Independent*. Retrieved from https://www.independent.co.uk/voices/artificial-intelligence-siri-cortana-alexa-music-pop-culture-robotics-a8406391.html

Deshpande, K. V., Pan, S., & Foulds, J. R. (2020). Mitigating Demographic Bias in AI-Based Resume Filtering. *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, 268–275. New York, NY, USA: Association for Computing Machinery. Scopus. https://doi.org/10.1145/3386392.3399569

Design Justice Network. (2018). Design Justice Network Principles. Retrieved from Design Justice Network website: https://designjustice.org/read-the-principles

D'Ignazio, C., & Klein, L. (2018). *Data Feminism. [Manuscript submitted for publication]*. MIT Open Press. Retrieved from https://mitpressonpubpub.mitpress.mit.edu/data-feminism

D'Ignazio, C., & Klein, L. (2020). *Data Feminism*. MIT Open Press. Retrieved from https://data-feminism.mitpress.mit.edu/

Dignum, V. (2019a). Ensuring Responsible AI in Practice. In V. Dignum (Ed.), *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way* (pp. 93–105). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-30371-6_6

Dignum, V. (2019b). Taking Responsibility. In V. Dignum (Ed.), *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way* (pp. 47–69). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-30371-6_4

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K. E., & Dugan, C. (2019). Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 275–285. Marina del Ray, CA, USA: ACM. https://doi.org/10.1145/3301275.3302310

Draude, C., Klumbyte, G., Lücking, P., & Treusch, P. (2019). Situated algorithms: A sociotechnical systemic approach to bias. *Online Information Review*, *44*(2), 325–342. https://doi.org/10.1108/OIR-10-2018-0332

Draude, C., Klumbyte, G., & Treusch, P. (2018). Re-Considering Bias: What Could Bringing Gender Studies and Computing Together Teach Us About Bias in Information Systems? *Proceedings of the International Workshop on Bias in Information, Algorithms, and Systems Co-Located with 13th International Conference on Transforming Digital Worlds (IConference 2018)*, 14–18. Sheffield, United Kingdom: CEUR-WS. Retrieved from http://ceur-ws.org/Vol-2103/#paper_3

Eitel-Porter, R. (2021). Beyond the promise: Implementing ethical AI. *AI and Ethics*, *1*(1), 73–80. https://doi.org/10.1007/s43681-020-00011-6

Etzkowitz, H., Kemelgor, C., Neuschatz, M., Uzzi, B., & Alonzo, J. (1994). The Paradox of Critical Mass for Women in Science. *Science*, *266*(5182), 51–54. Retrieved from https://www.jstor.org/stable/2884712

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. New York, New York: St. Martin’s Press.

European Commission. (2019). *Factsheet: Artificial Intelligence for Europe* (p. 2). Retrieved from https://ec.europa.eu/digital-single-market/en/news/factsheet-artificial-intelligence-europe

European Commission. (2021a, April 21). Europe fit for the Digital Age: Commission proposes new rules and actions for excellence and trust in Artificial Intelligence. Retrieved May 24, 2021, from The official website of the European Union website: https://ec.europa.eu/growth/content/europe-fit-digital-age-commission-proposes-new-rules-and-actions-excellence-and-trust_en

European Commission. (2021b, April 21). New rules for Artificial Intelligence – Q&As. Retrieved May 24, 2021, from European Commission—European Commission website: https://ec.europa.eu/commission/presscorner/detail/en/QANDA_21_1683

Feenberg, A. (1988). *The Bias Of Technology*. https://doi.org/10.1007/978-1-349-19275-5_12

Feenberg, A. (2003). What Is Philosophy of Technology? [Lecture for the Komaba undergraduates, June, 2003]. Retrieved from http://www.sfu.ca/~andrewf/komaba.htm

Feenberg, A. (2006). What Is Philosophy of Technology? In J. R. Dakers (Ed.), *Defining Technological Literacy: Towards an Epistemological Framework* (pp. 5–16). New York: Palgrave Macmillan US. https://doi.org/10.1057/9781403983053_2

Feng, P., & Feenberg, A. (2008). Thinking about Design: Critical Theory of Technology and the Design Process. In P. Kroes, P. E. Vermaas, A. Light, & S. A. Moore (Eds.), *Philosophy and Design: From Engineering to Architecture* (pp. 105–118). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-1-4020-6591-0_8

Forsythe, D. E. (2001). *Studying Those Who Study Us: An Anthropologist in the World of Artificial Intelligence*. Stanford: Stanford University Press. Retrieved from http://www.sup.org/books/title/?id=1342

Gaucher, D., Friesen, J., & Kay, A. C. (2011). Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology*, *101*(1), 109–128. https://doi.org/10.1037/a0022530

Gender. (2020). *Merriam-Webster.Com*. Retrieved from https://www.merriam-webster.com/dictionary/gender

Giannoumis, A., & Bui, C. (2019). *Gender Bias in AI [Presentation Slides]*. Unpublished research presented at the 1st KAIST International Conference on AI Fairness: AI and Gender, Daejeon, South Korea. Retrieved from https://docs.google.com/presentation/d/19ECDdGzFpxDtARJx6dhGldnsGRX5-sJoUus4DOiDFlI/edit?usp=sharing

Global Gender Gap Report 2020. (2019, December 16). Retrieved January 26, 2021, from World Economic Forum website: https://www.weforum.org/reports/gender-gap-2020-report-100-years-pay-equality/

Gordon-Murnane, L. (2018). Ethical, Explainable Artificial Intelligence: Bias and Principles. *Online Searcher*, *42*(2), 22–44. Education Research Complete. Retrieved from Education Research Complete.

Harding, S. (1992). Rethinking Standpoing Epistemology: What is “Strong Objectivity”? *The Centennial Review*, *36*(3), 437–470. JSTOR. Retrieved from http://www.jstor.org/stable/23739232

Harding, S. (1995). “Strong objectivity”: A response to the new objectivity question. *Synthese*, *104*(3), 331–349. https://doi.org/10.1007/BF01064504

Harding, S. (Ed.). (2004). *The feminist standpoint theory reader: Intellectual and political controversies*. New York: Routledge.

Harkness, T. (2020, August 18). How Ofqual failed the algorithm test. *UnHerd*. Retrieved from https://unherd.com/2020/08/how-ofqual-failed-the-algorithm-test/?tl_inbound=1&tl_groups[0]=18743&tl_period_type=3

Harwell, D. (2018, December 30). Fake-porn videos are being weaponized to harass and humiliate women: ‘Everybody is a potential target’—The Washington Post. Retrieved January 22, 2021, from The Washington Post website: https://www.washingtonpost.com/technology/2018/12/30/fake-porn-videos-are-being-weaponized-harass-humiliate-women-everybody-is-potential-target/

Hunter, A. P., Sheppard, L. R., Karlén, R., & Balieiro, L. (2018). *Adoption of Artificial Intelligence* (pp. 24–34). Center for Strategic and International Studies (CSIS). JSTOR. https://doi.org/10.2307/resrep22492.7

Huston, M. (2018, September 7). 12 Common Biases That Affect How We Make Everyday Decisions [Magazine]. Retrieved May 7, 2021, from Psychology Today website: https://www.psychologytoday.com/us/blog/thoughts-thinking/201809/12-common-biases-affect-how-we-make-everyday-decisions

IBM. (2019). *Everyday Ethics for Artificial Intelligence*. IBM. Retrieved from IBM website: https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf

IBM Developer Staff. (2020, March 9). AI Fairness 360. Retrieved from https://developer.ibm.com/technologies/artificial-intelligence/projects/ai-fairness-360/

McKinsey & Company. (2019, December 4). “Invisible Women” wins the 2019 Business Book of the Year Award [Blog]. Retrieved April 14, 2021, from McKinsey & Company website: https://www.mckinsey.com/about-us/new-at-mckinsey-blog/2019-business-book-of-the-year-award

Jegadeesan, M. (2020, January 5). *Adversarial Demotion of Bias in Natural Language Generation*. 2 pages. New York, NY, USA: ACM. https://doi.org/10.1145/3371158.3371229

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Johnson, K. (2021, January 20). Google targets AI ethics lead Margaret Mitchell after firing Timnit Gebru. *VentureBeat*. Retrieved from https://venturebeat.com/2021/01/20/google-targets-ai-ethics-lead-margaret-mitchell-after-firing-timnit-gebru/

Jones, M. L. (2018). How We Became Instrumentalists (Again). *Historical Studies in the Natural Sciences*, *48*(5), 673–684. https://doi.org/10.1525/hsns.2018.48.5.673

Jonhson, K. (2020, December 4). AI ethics pioneer's exit from Google involved research into risks and inequality in large language models. Retrieved February 2, 2021, from VentureBeat website: https://venturebeat.com/2020/12/03/ai-ethics-pioneers-exit-from-google-involved-research-into-risks-and-inequality-in-large-language-models/

Jule, A. (2014). Gender Theory. In A. C. Michalos (Ed.), *Encyclopedia of Quality of Life and Well-Being Research* (pp. 2464–2466). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-0753-5_1137

Kakarmath, S., Esteva, A., Arnaout, R., Harvey, H., Kumar, S., Muse, E., … Kvedar, J. (2020). Best practices for authors of healthcare-related artificial intelligence manuscripts. *Npj Digital Medicine*, *3*(1), 1–3. https://doi.org/10.1038/s41746-020-00336-w

Kantayya, S. (2020). *Coded Bias*. USA. Retrieved from https://www.codedbias.com/about

Kay, M., Matuszek, C., & Munson, S. A. (2015). Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 3819–3828. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2702123.2702520

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *35th International Conference on Machine Learning, ICML 2018*, *6*, 4008–4016. International Machine Learning Society (IMLS). Scopus. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-85057263888&partnerID=40&md5=2f45a98ed96b2e1b4e76d08536e9f718

Kim, D. H. (1999). Introduction to Systems Thinking. *Pegasus Communications, Inc.*, 21. Retrieved from https://thesystemsthinker.com/wp-content/uploads/2016/03/Introduction-to-Systems-Thinking-IMS013Epk.pdf

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., … Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 201915768. https://doi.org/10.1073/pnas.1915768117

Krause, H. (2019, November 7). Who is the Head of Your Household? Retrieved February 5, 2021, from We All Count website: https://weallcount.com/2019/11/07/who-is-the-head-of-your-household/

Kristiansen, E. (2020, October 13). Hvorfor så få kvinner i IT-bransjen? [News]. Retrieved May 12, 2021, from Digi.no website: https://www.digi.no/artikler/debatt-hvorfor-sa-fa-kvinner-i-it-bransjen/500823

Lambrecht, A., & Tucker, C. (2019). Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Management Science*, *65*(7), 2966–2981. https://doi.org/10.1287/mnsc.2018.3093

Lazovich, T. (2020). Does Deep Learning Have Politics? *Resistance AI Workshop @NeurIPS 2020*, 4. Virtual: Resistance AI Workshop @NeurIPS 2020. Retrieved from https://sites.google.com/view/resistance-ai-neurips-20/accepted-papers-and-media

Leavy, S. (2018). Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. *Proceedings - International Conference on Software Engineering*, 14–16. Gothenburg, Sweden: IEEE Computer Society. Scopus. https://doi.org/10.1145/3195570.3195580

Leicht-Deobald, U., Busch, T., Schank, C., Weibel, A., Schafheitle, S., Wildhaber, I., & Kasper, G. (2019). The Challenges of Algorithm-Based HR Decision-Making for Personal Integrity. *Journal of Business Ethics*, *160*(2), 377–392. https://doi.org/10.1007/s10551-019-04204-w

LEXIN. (2019). LEXIN Dictionary. In *LEXIN English-Bokmål Dictionary*. Oslo Metropolitan University. Retrieved from https://lexin.oslomet.no/#/findwords/message.bokmal-english

Maxwell, J. A. (2012). Conceptual Framework: What do you think is going on? In *Qualitative Research Design: An Interactive Approach (3rd Edition)* (pp. 39–72). SAGE Publications, Inc.

Meadows, D. (1999). Leverage points: Places to Intervene in a System. *The Sustainability Institute*. Retrieved from http://www.donellameadows.org/wp-content/userfiles/Leverage_Points.pdf

Melendez, S. (2018, August 9). Uber driver troubles raise concerns about transgender face recognition. Retrieved January 27, 2021, from Fast Company website: https://www.fastcompany.com/90216258/uber-face-recognition-tool-has-locked-out-some-transgender-drivers

MIT Institutional Research. (n.d.). Diversity Dashboard. Retrieved June 12, 2021, from MIT Institutional Research website: https://ir.mit.edu/diversity-dashboard

MIT Media Lab. (2018). Norman, World's first psychopath AI. Retrieved April 15, 2020, from http://norman-ai.mit.edu/

Mohla, S., Bagh, B., & Guha, A. (2021). A Material Lens to Investigate the Gendered Impact of the AI Industry. *IJCAI 2021 Workshop on AI for Social Good*. Retrieved from https://crcs.seas.harvard.edu/files/crcs/files/ai4sg-21_paper_8.pdf

Motapanyane, J.-M. (2014). Feminism, an Overview. In A. C. Michalos (Ed.), *Encyclopedia of Quality of Life and Well-Being Research* (pp. 2246–2250). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-0753-5_1035

Myers, M. D. (Ed.). (1997). Qualitative Research in Information Systems. *Association for Information*

Systems (AISWorld) Section on Qualitative Research in Information Systems. Originally published in MISQ Discovery, June 1997. Retrieved from https://www.qual.auckland.ac.nz/

Myers, M. D., & Klein, H. (2011). A Set of Principles for Conducting Critical Research in Information Systems. *MIS Q.*, *35*, 17–36. https://doi.org/10.2307/23043487

Myers, M. D., & Newman, M. (2007). The qualitative interview in IS research: Examining the craft. *Information and Organization*, *17*(1), 2–26. https://doi.org/10.1016/j.infoandorg.2006.11.001

Naughton, J. (2020, September 6). From viral conspiracies to exam fiascos, algorithms come with serious side effects. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2020/sep/06/from-viral-conspiracies-to-exam-fiascos-algorithms-come-with-serious-side-effects

Nissenbaum, H. (1996). Accountability in a computerized society. *Science and Engineering Ethics*, *2*(1), 25–42. https://doi.org/10.1007/BF02639315

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York: New York University Press.

NOU 2019: 5. (2019). *Ny forvaltningslov—Lov om saksbehandlingen i offentlig forvaltning (forvaltningsloven)* (NOU No. 5). Oslo: Norges offentlige utredninger. Retrieved from Norges offentlige utredninger website: https://www.regjeringen.no/no/dokumenter/nou-2019-5/id2632006/

Nova. (2021). The global top talent network. Retrieved May 6, 2021, from Nova Talent website: https://novatalent.com/

NVivo. (2021). Learn More About Data Analysis Software. Retrieved April 20, 2021, from https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/about/nvivo

ODA. (n.d.). ODA er Nordens ledende møteplass for kvinner i tech. Vår visjon: Lead the Change! Retrieved May 6, 2021, from ODA-Nettverk website: http://odanettverk.no

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Broadway Books. Retrieved from http://governance40.com/wp-content/uploads/2019/03/Weapons-of-Math-Destruction-Cathy-ONeil.pdf

Otterbacher, J., Checco, A., Demartini, G., & Clough, P. (2018). Investigating User Perception of Gender Bias in Image Search: The Role of Sexism. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 933–936. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3209978.3210094

Papenfuss, M. (2017, December 15). Woman In China Says Colleague's Face Was Able To Unlock Her iPhone X. *HuffPost*. Retrieved from https://www.huffpost.com/entry/iphone-face-recognition-double_n_5a332cbce4b0ff955ad17d50

Parsheera, S. (2018, November 26). *A gendered perspective on artificial intelligence*. 1–7.

https://doi.org/10.23919/ITU-WT.2018.8597618

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Cambridge: Harvard University Press.

Perez, C. C. (2019). *Invisible Women: Exposing Data Bias in a World Designed for Men*. London: Chatto & Windus.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). *On Fairness and Calibration*. 10. Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/7151-on-fairness-and-calibration.pdf

PRE-CRIME. (n.d.). Retrieved February 5, 2021, from PRE-CRIME website: http://precrime-film.com

Pronin, E., Lin, D. Y., & Ross, L. (2002). The Bias Blind Spot: Perceptions of Bias in Self Versus Others. *Personality and Social Psychology Bulletin*, *28*(3), 369–381. https://doi.org/10.1177/0146167202286008

Quan-Haase, A. (2013). Theoretical perspectives on technology. In *Technology and society: Social networks, power and inequality* (pp. 42–61). Don Mills Ont Oxford: Oxford University Press.

Raji, I. D., & Buolamwini, J. (2019). *Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products*. 429–435. Honolulu, HI, USA: Association for Computing Machinery. https://doi.org/10.1145/3306618.3314244

Rémy, P. (2021). *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Retrieved from https://github.com/philipperemy/timit

Richardson, R., Schultz, J., & Crawford, K. (2019). *Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice* (SSRN Scholarly Paper No. ID 3333423). Rochester, NY: Social Science Research Network. Retrieved from Social Science Research Network website: https://papers.ssrn.com/abstract=3333423

Robson, C. (2002). *Real World Research: A Resource for Social Scientists and Practitioner-Researchers* (2nd ed.). Oxford: Blackwell.

Schaffer, C. (1993). Overfitting avoidance as bias. *Machine Learning*, *10*(2), 153–178. https://doi.org/10.1007/BF00993504

Schei, T. H. (2020). *IHuman*. Norway. Retrieved from https://tv.nrk.no/program/KOID75003817

Schwartz-Ziv, M. (2017). Gender and Board Activeness: The Role of a Critical Mass. *Journal of Financial and Quantitative Analysis*, *52*(2), 751–780. https://doi.org/10.1017/S0022109017000059

Sharkey, N. (2014). Towards a principle for the human supervisory control of robot weapons. *Politica & Societa*, (2), 305–324. https://doi.org/10.4476/77105

Shilton, K. (2013). Values Levers: Building Ethics into Design. *Science, Technology, & Human Values*, *38*(3), 374–397. https://doi.org/10.1177/0162243912436985

Shilton, K. (2014). This is an Intervention: Foregrounding and Operationalizing Ethics During Technology Design. In K. D. Pimple (Ed.), *Emerging Pervasive Information and Communication Technologies (PICT): Ethical Challenges, Opportunities and Safeguards* (pp. 177–192). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-6833-8_9

Specia, M. (2019, May 22). Siri and Alexa Reinforce Gender Bias, U.N. Finds. *The New York Times*. Retrieved from https://www.nytimes.com/2019/05/22/world/siri-alexa-ai-gender-bias.html

Spotlight—Coded Bias Documentary. (n.d.). Retrieved April 14, 2021, from https://www.ajl.org/spotlight-documentary-coded-bias

Språkbehandling og kunstig intelligens [Video file]. (2020). Virtual: Tekna. Retrieved from https://www.tekna.no/fag-og-nettverk/IKT/ikt-bloggen/sprakbehandling-og-kunstig-intelligens/

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*(56), 1929–1958. Retrieved from http://jmlr.org/papers/v15/srivastava14a.html

Stumpf, S., Peters, A., Bardzell, S., Burnett, M., Busse, D., Cauchard, J., & Churchill, E. (2020). Gender-Inclusive HCI Research and Design: A Conceptual Review. *Foundations and Trends® in Human–Computer Interaction*, *13*(1), 1–69. https://doi.org/10.1561/1100000056

Tatman, R. (2017). Gender and Dialect Bias in YouTube's Automatic Captions. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 53–59. Valencia, Spain: Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1606

Taylor, S. J., Bogdan, R., & DeVault, M. L. (2016). *Introduction to qualitative research methods: A guidebook and resource* (4th ed.). Hoboken, New Jersey: John Wiley & Sons, Inc. Retrieved from http://www.elfhs.ssru.ac.th/pokkrong_ma/pluginfile.php/50/block_html/content/%5bTaylor,_Steven;_Bogdan,_Robert;_DeVault,_Marjorie(b-ok.org).pdf

The Future of Life Institute. (2017). Asimolar AI Principles. Retrieved April 20, 2021, from https://futureoflife.org/ai-principles/

Thelwall, M. (2018). Gender bias in machine learning for sentiment analysis. *Online Information Review*, *42*(3), 343–354. https://doi.org/10.1108/oir-05-2017-0153

Thompson, C. (2019, February 13). The Secret History of Women in Coding [News]. Retrieved May 12, 2021, from The New York Times website: https://www.nytimes.com/2019/02/13/magazine/women-coding-computer-programming.html

Tikkanen, A. (n.d.). Titanic. In *Encyclopedia Britannica*. Retrieved from https://www.britannica.com/topic/Titanic

Timcke, S. (2020). *Algorithms and the Critical Theory of Technology* (SSRN Scholarly Paper No. ID 3551467). Rochester, NY: Social Science Research Network.

https://doi.org/10.2139/ssrn.3551467

United Nations University & EQUALS. (2019). *Taking Stock: Data and Evidence on Gender Equality in Digital Access, Skills, and Leadership*. Macau: EQUALS Global Partnership. Retrieved from EQUALS Global Partnership website: https://2b37021f-0f4a-4640-8352-0a3c1b7c2aab.filesusr.com/ugd/04bfff_145a18e6425e47a1b90d0440f7476d0f.pdf

Verbeek, P.-P. (2008). Morality in Design: Design Ethics and the Morality of Technological Artifacts. In P. Kroes, P. E. Vermaas, A. Light, & S. A. Moore (Eds.), *Philosophy and Design: From Engineering to Architecture* (pp. 91–103). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-1-4020-6591-0_7

Vincent, J. (2016, March 24). Twitter taught Microsoft’s friendly AI chatbot to be a racist asshole in less than a day. Retrieved February 11, 2021, from The Verge website: https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

Wachter, S., Mittelstadt, B., & Russell, C. (2021). *Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law* (SSRN Scholarly Paper No. ID 3792772). Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.3792772

Wagner, B. (2018). Ethics as an Escape from Regulation.: From “Ethics-Washing” to Ethics-Shopping? In E. Bayamlioglu, I. Baraliuc, L. Janssens, & M. Hildebrandt (Eds.), *Being Profiled* (pp. 84–89). Amsterdam University Press. https://doi.org/10.2307/j.ctvhrd092.18

Wakefield, J. (2018, June 1). Are you scared yet? Meet Norman, the psychopathic AI. *BBC*. Retrieved from https://www.bbc.com/news/technology-44040008

We All Count. (n.d.-a). A Matter of Life and Death. Retrieved February 5, 2021, from We All Count website: https://weallcount.com/home-legacy/

We All Count. (n.d.-b). Data Equity Framework. Retrieved February 5, 2021, from We All Count website: https://weallcount.com/the-data-process/

Wellner, G., & Rothman, T. (2020). Feminist AI: Can We Expect Our AI Systems to Become Feminist? *Philosophy & Technology*, *33*(2), 191–205. https://doi.org/10.1007/s13347-019-00352-z

West, M., Kraut, R., & Chew, H. E. (2019). *I’d Blush If I Could*. Geneva: EQUALS and The United Nations Educational, Scientific and Cultural Organization. Retrieved from EQUALS and The United Nations Educational, Scientific and Cultural Organization website: UNESCO website https://en.unesco.org/Id-blush-if-I-could

West, S. M., Whittaker, M., & Crawford, K. (2019). *Discriminating Systems: Gender, Race and Power in AI* (p. 33). AI Now Institute. Retrieved from AI Now Institute website: https://ainowinstitute.org/discriminatingsystems.html

Winner, L. (1980). Do Artifacts Have Politics? *Daedalus*, *109*(1), 121–136. JSTOR. Retrieved from http://www.jstor.org/stable/20024652

World Health Organization. (2011). Gender, equity and human rights. Retrieved May 10, 2021, from

World Health Organization website: https://www.who.int/gender-equity-rights/knowledge/glossary/en/

XMind. (2021). XMind—Full-featured mind mapping and brainstorming tool. Retrieved April 20, 2021, from XMind website: https://www.xmind.net

Yin, R. K. (2009). *Case Study Research: Design and Methods* (4th ed.). Thousand Oaks, Calif: Sage.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). *Learning Fair Representations* (D. Sanjoy & M. David, Eds.). Proceedings of Machine Learning Research: PMLR. Retrieved from http://proceedings.mlr.press

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340. New Orleans LA USA: ACM. https://doi.org/10.1145/3278721.3278779

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2017). Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2979–2989. Copenhagen, Denmark: Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1323

Zook, M., Barocas, S., Boyd, D., Crawford, K., Keller, E., Gangadharan, S. P., … Pasquale, F. (2017). Ten simple rules for responsible big data research. *PLOS Computational Biology*, *13*(3), e1005399. https://doi.org/10.1371/journal.pcbi.1005399

Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist—It's time to make it fair. *Nature*, *559*(7714), 324–326. https://doi.org/10.1038/d41586-018-05707-8

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. New York: PublicAffairs.

Table and Figure References

References for figures and tables not listed in the References list are below.

Academy for Systems Change. (n.d.). Systems Thinking Resources. Retrieved April 23, 2021, from The Academy for Systems Change website: http://donellameadows.org/systems-thinking-resources/

Feenberg, A. (2006). *Defining Technological Literacy Towards an Epistemological Framework.* New York, NY: Palgrave Macmillan US. Retrieved from https://link.springer.com/chapter/10.1057/9781403983053_2

Appendices

Appendix A - Perspectives in a Titanic Analogy

This analogy was used during the analysis of the data to get an understanding of the different patterns and groups that emerged from the data.

The history of Titanic is widely known. It was a massive ship on the way from England to the US in 1912 (Tikkanen, n.d.). It was the biggest ship ever made and it was said to be unsinkable. However, it hit an iceberg that the crew was unable to see until it was too late. Because of the nature and weight of icebergs only 10% of their total size is visible from the surface. The crash between the ship and the iceberg led to Titanic sinking into the sea leaving an estimated number of 1 500 victims (Tikkanen, n.d.).

This history is useful to have in mind as the characteristics of the different groups found in the data are outlined. After analyzing the data approximately 4 different groups of practitioners were identified:

1. The Guests
2. The Big Crew
3. The Little Crew
4. The Lighthouse Worker & the Helicopters
5. The Divers

These groups are not clearly demarcated and some of the practitioners have traits from several groups. There is no doubt among all of these groups that crashing the ship, i.e. bias in AI systems, is bad. They are also all open to the opportunity that there are biases present that they are unable to see. However, they address gender bias and unconscious bias in AI to varying degrees. In this analogy, the ship represents the AI system. The journey of the ship is a metaphor for the system's development and use. The iceberg represents biases that can show up along the way and wreck the use or the development of the system. As outlined in the background chapter and similar to the Titanic, this could lead to fatal consequences as serious as life and death.

The Guests: Techno-blind

The Guests are characterized by their somewhat carefree outlook and attitude. They assume that everything is fine and they assume that the ship will not crash. When you ask them whether they think the ship will crash, they look at the horizon with their binoculars and since they cannot see any imminent dangers they reply: "Can't see any reason why. But I might be wrong". They also have great trust in the science and robustness of the biggest ship ever made. Their trust in the science and engineers who created the ship blinds them and makes it harder for them to comprehend that the way the ship is designed might be a weakness.

For instance, PC4 understands the importance of gender balance within the research subjects. However, her concern is more about the statistical correctness of the research rather than the potential human impact. A symptom of this technical view is how she addresses inequality issues like racism with technical words to describe it.

Informants in this category: AR2, PC4

The Big Crew: Techno-ignorant and Passive

The Big Crew work on the big ship and are characterized by that they do understand that they have some responsibility to keep the ship running. However, since the coal diggers and deck workers are so separated they only see their responsibility as the tasks that are right in front of them

They have the ability to see the entirety of the iceberg because they use an underwater binocular to look for dangers, but because the range of the binoculars are so short they are unable to see the dangers that are far ahead.

They check the sight whenever they feel like it. Most of the time when they look in the underwater binoculars they don't see any dangers. They might then assume that there are no dangers ahead or not check very often since they usually do not see any dangers. The Big Crew members were not able to think of many ethical concerns when they were asked about it.

Informants in this category: AR1, AR4

The Little Crew: Techno-ignorant, but Active

The Little Crew are an extended ship's crew that do little outings on a rubber boat to check for dangers. They have similarities to The Big Crew but they have an increased level of awareness and more ethical practices. These "outings" are signified as that the teams in The Little Crew have ethical discussions and ethical interventions.

SU3 decided to circumvent the entire citation system when they developed an AI system for finding research articles. They believed that embedding the citation system in the AI would have exacerbated an already biased system. SU2 and SU3 of The Little Crew lack the resources for diving since they are startups. Because they make occasional boat trip excursions, they are better able to understand the problem of icebergs if they were to encounter one, compared to the Big Crew. However, the reach of the rubber boat is still short and their understanding of the icebergs are limited compared to the divers. Although there are several things that they do right and understand, they lack control points, regular processes, and testing practices for continually checking for biases.

Informants in this category: SU3, PC1, SU2

The Lighthouse Worker: The Techno-Optimist

The Lighthouse Worker is a worker with good intentions. They are aware that icebergs are dangerous and they want to prevent other ships from crashing into icebergs. So they set up lighthouses or flags on icebergs that they find in order to warn ships. However, they do not know that they only see the tip of the iceberg and is unable to convey this information to the ships. They have a simplified understanding of the issue and believe that the main problem is that the ships cannot see the icebergs. So that even if a ship does see their light and tries to avoid the iceberg, without the knowledge about the scope of the unseen ice, they might still crash into the parts that are below the surface.

Informants in this category: SU1

The Helicopter: Techno-Capitalist

The Helicopter Crew's job is to fly over the ocean to look for risks and dangers for the ship. But they are not looking for biases, they are looking for business risks. They have some practices that are similar to ethical practices, however their goals are not to increase ethics, but to do analysis of technical risks and potential customer loss. They are looking for things that can cause scandals, loss of profit and reputation.

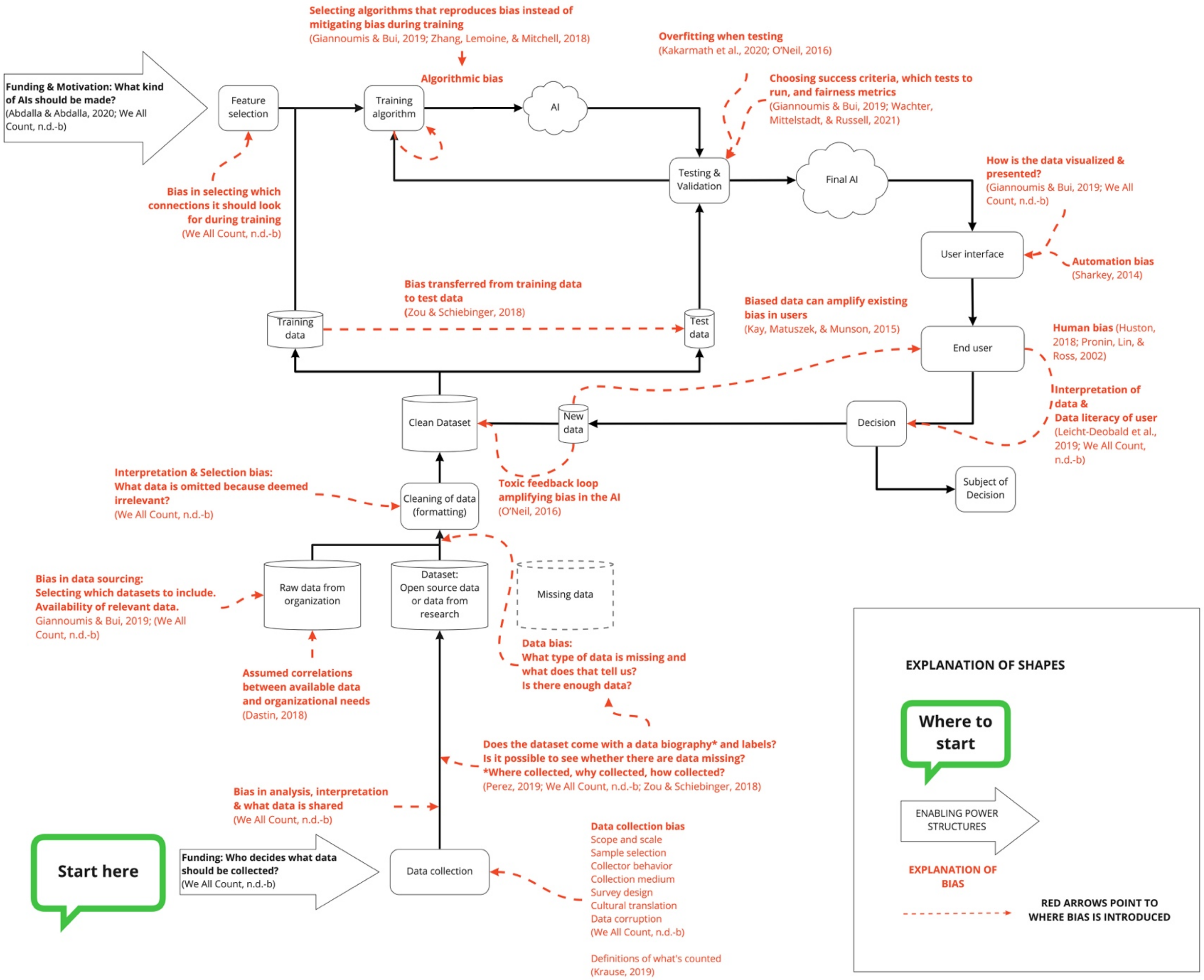
Informants in this category: PC3, GI1

The Divers: Techno-Ethicists

The divers go through the trouble of diving below the surface because they understand the necessity of scoping the size of the iceberg in order for the ship to safely pass through that area. But this requires the resources of diving equipment that perhaps not everyone can afford. All three divers, in addition to PC3, were concerned with definitions of fairness, and pointed to this as one of the big causes. Meant structures of society were one of the causes.

Informants in this category: GI2, PC2, AR3

Appendix B - Figure of Entry Points of Bias in AI Systems, Detailed Version


Funding & Motivation: What kind of AIs should be made?
(Abdalla & Abdalla, 2020; We All Count, n.d.-b)
Feature selection
Training algorithm
AI
Testing & Validation
Final AI
User interface
End user
Decision
Subject of Decision
Training data
Test data
New data
Clean Dataset
Cleaning of data (formatting)
Raw data from organization
Dataset: Open source data or data from research
Missing data
Data collection
Start here
Funding: Who decides what data should be collected?
(We All Count, n.d.-b)
EXPLANATION OF SHAPES
Where to start
ENABLING POWER STRUCTURES
EXPLANATION OF BIAS
RED ARROWS POINT TO WHERE BIAS IS INTRODUCED

Selecting algorithms that reproduces bias instead of mitigating bias during training
(Giannoumis & Bui, 2019; Zhang, Lemoine, & Mitchell, 2018)

Algorithmic bias

Overfitting when testing
(Kakarmath et al., 2020; O'Neil, 2016)

Choosing success criteria, which tests to run, and fairness metrics
(Giannoumis & Bui, 2019; Wachter, Mittelstadt, & Russell, 2021)

How is the data visualized & presented?
(Giannoumis & Bui, 2019; We All Count, n.d.-b)

Automation bias
(Sharkey, 2014)

Bias in selecting which connections it should look for during training
(We All Count, n.d.-b)

Bias transferred from training data to test data
(Zou & Schiebinger, 2018)

Biased data can amplify existing bias in users
(Kay, Matuszek, & Munson, 2015)

Human bias (Huston, 2018; Pronin, Lin, & Ross, 2002)

Interpretation of data &
Data literacy of user
(Leicht-Deobald et al., 2019; We All Count, n.d.-b)

Interpretation & Selection bias:
What data is omitted because deemed irrelevant?
(We All Count, n.d.-b)

Toxic feedback loop amplifying bias in the AI
(O'Neil, 2016)

Bias in data sourcing:
Selecting which datasets to include.
Availability of relevant data.
Giannoumis & Bui, 2019; (We All Count, n.d.-b)

Data bias:
What type of data is missing and what does that tell us?
Is there enough data?

Assumed correlations between available data and organizational needs
(Dastin, 2018)

Does the dataset come with a data biography* and labels?
Is it possible to see whether there are data missing?
***Where collected, why collected, how collected?**
(Perez, 2019; We All Count, n.d.-b; Zou & Schiebinger, 2018)

Bias in analysis, interpretation & what data is shared
(We All Count, n.d.-b)

Data collection bias
Scope and scale
Sample selection
Collector behavior
Collection medium
Survey design
Cultural translation
Data corruption
(We All Count, n.d.-b)

Definitions of what's counted
(Krause, 2019)

Appendix C - Interview Guide

1. Questions about their **experience** and expertise.

    a. Can you tell me about your research in AI? / What is your experience working with AI?

2. Questions about their organization and **development process**

    a. Can you describe the AI development process?

        i. If in person: Can you draw a model of how you work with AI?

    b. Are there any ethical concerns you need to address in your work and can you elaborate on that?

3. Knowledge / awareness of **bias**

    a. If bias is mentioned, go to gender bias.

    b. If bias is not mentioned during the previous question:
    There is a lot of literature suggesting that bias is an ethical issue in the field of AI. Has the issue of bias been discussed in your organization?

4. Knowledge/awareness of **gender bias** and ethical processes
There have been articles in the news on issues such as how facial or voice recognition does not work as well for women [choose example close to their work].

    a. If bias has been mentioned:
    Has gender bias been discussed as one of the biases in the team or company?

    b. How or in what ways is gender bias taken into account in your work?

5. **Causes** of gender bias

    a. What do you think might be the cause of gender bias in AI?
    (What component in the system?)

b. If they have mentioned gender bias having been discussed before:
   What is your experience with where gender bias comes into an AI system?

6. Ethical guidelines

   a. Several organizations and companies have published ethical guidelines for developing AI systems. Does your company use one and why or why not?

   b. What do you think would be a good way to motivate your company/organization to try to solve issues of gender bias in AI?

Other:

- Can I have access to your research proposal? *
- Ask to have them send ethics guidelines they are following/relevant for their company *
- Can you reflect on the role of gender bias in your research project/work?

* This might be too time-consuming to process

Appendix D - Consent Form and Information Letter Ver. 1

Are you interested in taking part in the research project

”Ethics in Artificial Intelligence”?

This is an inquiry about participation in a research project where the main purpose is to investigate ethics practices and knowledge about ethics in the field of AI. In this letter we will give you information about the purpose of the project and what your participation will involve.

Purpose of the project
The master’s thesis project aims to investigate current ethics practices and collect knowledge on good practices for ethics in Artificial Intelligence (AI). Data collected will be used to map the challenges of ethics in AI into a visual model. Data will also be used to outline recommendations for good practices. This project aims to conduct 5-20 interviews. Results of this thesis might be used to publish scientific articles.

Who is responsible for the research project?
The University of Oslo is the institution responsible for the project.

Why are you being asked to participate?
You have indicated that you work in the field of AI as a researcher, practitioner, leader, developer, or other. Someone you know might have recommended you as a participant to the project, and/or you have been contacted through a publicly available online profile. Approximately 10-40 people in Norway and other parts of the world have been/will be asked to participate in this project.

What does participation involve for you?
If you choose to take part in the project, participation will involve an interview. The interview will last approximately 60 minutes. Your answers will be saved as a sound recording.

The interview includes questions about:

- Your professional experience and expertise in AI and ethics.
- Ethics practices and processes.
- The development and creation of AI systems.

Participation is voluntary
Participation in the project is voluntary. If you choose 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 choose 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 purposes specified in this information letter. We will process your personal data confidentially and in accordance with data protection legislation (the General Data Protection Regulation and Personal Data Act).

- Only the student and supervisor will have access to any data collected.
- The student will replace your name and contact details with a code. The list of names, contact details, and respective codes will be password protected and stored separately from the recordings and transcripts.
- Recordings will be deleted as soon as a transcript has been made of the recording.
- Recordings and transcripts will be stored on a secure and locked cloud service at the University of Oslo.

Anonymized and aggregated statistics on occupation and demography might be included in the thesis (e.g. sector, generic position, gender). Any quotes included in the published thesis will be anonymized and participants will not be recognizable.

What will happen to your personal data at the end of the research project?

The project is scheduled to end in July 2021. Personal data will be deleted at the end of the project.

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 Data Protection Officer or 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 the University of Oslo, NSD – The Norwegian Centre for Research Data AS has assessed that the processing of personal data in this project is in accordance with data protection legislation.

Where can I find out more?

If you have questions about the project, or want to exercise your rights, contact:

- The University of Oslo via Cathrine Bui (ckbui@ifi.uio.no) or Maja Van der Velden (majava@ifi.uio.no).
- Our Data Protection Officer: Roger Markgraf-Bye (personvernombud@uio.no).
- NSD – The Norwegian Centre for Research Data AS, by email: (personverntjenester@nsd.no) or by telephone: +47 55 58 21 17. Reference number for the project: 816757.

Best regards,

Supervisor Maja Van der Velden

Student Cathrine Bui

I have received and understood information about the project “Ethics in AI” and have been given the opportunity to ask questions. I give consent:

□ To participate in an interview
□ To be recorded during the interview

I give consent for my personal data to be processed until the end date of the project, approx. July 2021.

(Name in upper case letters) (Signature) (Date)

Appendix E -
Consent Form and Information Letter Ver. 2

Main changes from version 1 of the consent form and information letter are highlighted in yellow.

9 Are you interested in taking part in the research project *"Ethics in Artificial Intelligence"*?

This is an inquiry about participation in a research project where the main purpose is to investigate ethics practices and knowledge about ethics in the field of AI. In this letter we will give you information about the purpose of the project and what your participation will involve.

Purpose of the project
The master's thesis project aims to investigate current ethics practices and collect knowledge on good practices for ethics in Artificial Intelligence (AI). Data collected will be used to map the challenges of ethics in AI into a visual model. Data will also be used to outline recommendations for good practices. This project aims to conduct 5-20 interviews. Results of this thesis might be used to publish scientific articles or additional research projects.

Who is responsible for the research project?
The University of Oslo is the institution responsible for the project.

Why are you being asked to participate?
You have indicated that you work in the field of AI as a researcher, practitioner, leader, developer, or other. Someone you know might have recommended you as a participant to the project, and/or you have been contacted through a publicly available online profile. Approximately 10-40 people in Norway have been/will be asked to participate in this project.

What does participation involve for you?
If you choose to take part in the project, participation will involve an interview. The interview will last approximately 60 minutes. Your answers will be saved as a sound recording.

The interview includes questions about:

- Your professional experience and expertise in AI and ethics.
- Ethics practices and processes.
- The development and creation of AI systems.

The follow-up interview will last approximately 5 minutes and will be over the phone. This call will not be recorded. This interview will be about *how many* marginalized identities you identify with and may therefore collect information about topics such as health and sexual orientation. You will **not** be asked to state *which* identities you identify with, apart from gender.

Participation is voluntary

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

10

11 **Your personal privacy – how we will store and use your personal data**

We will only use your personal data for the purposes specified in this information letter. We will process your personal data confidentially and in accordance with data protection legislation (the General Data Protection Regulation and Personal Data Act).

- The student and supervisor will have access to any data collected.
- The student will replace your name and contact details with an ID code. The list of names, contact details and respective codes will be password protected and stored separately from the recordings and transcripts.
- Recordings and transcripts will be stored on a secure and locked cloud service at the University of Oslo.
- Transcripts and de-identified research notes may be safely archived in a different location.
- Data from the follow-up interview will be stored on an encrypted USB stick and stored separately from the list of names.

What will happen to your personal data at the end of the research project?

The project is scheduled to end in December 2021. The list of names and connecting ID codes of participants and recordings will be deleted at project end. Only fully anonymised research data will be archived. Archived data will be stored for an unspecified length of time.

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 Data Protection Officer or 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 the University of Oslo, NSD – The Norwegian Centre for Research Data AS has assessed that the processing of personal data in this project is in accordance with data protection legislation.

Where can I find out more?

If you have questions about the project, or want to exercise your rights, contact:

- The University of Oslo via Cathrine Bui (ckbui@ifi.uio.no) or Maja Van der Velden (majava@ifi.uio.no).
- Our Data Protection Officer: Roger Markgraf-Bye (personvernombud@uio.no).

- NSD – The Norwegian Centre for Research Data AS, by email: (personverntjenester@nsd.no) or by telephone: +47 55 58 21 17. Reference number for the project: 816757.

Best regards,

Supervisor Maja Van der Velden
Student Cathrine Bui

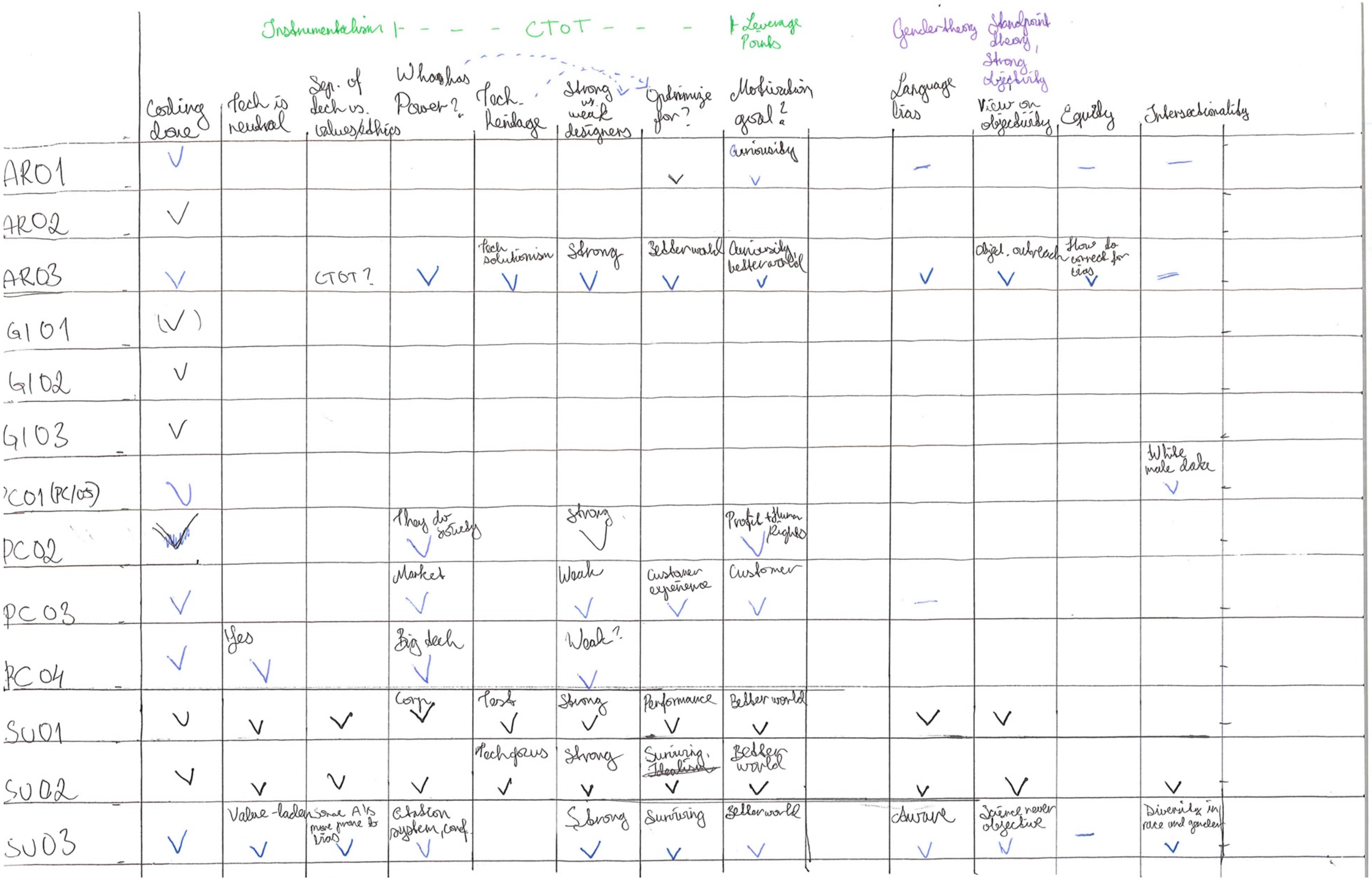
I have received and understood information about the project "Ethics in AI" and have been given the opportunity to ask questions. I give consent:

□ To participate in interviews.
□ To be recorded during the first interview.

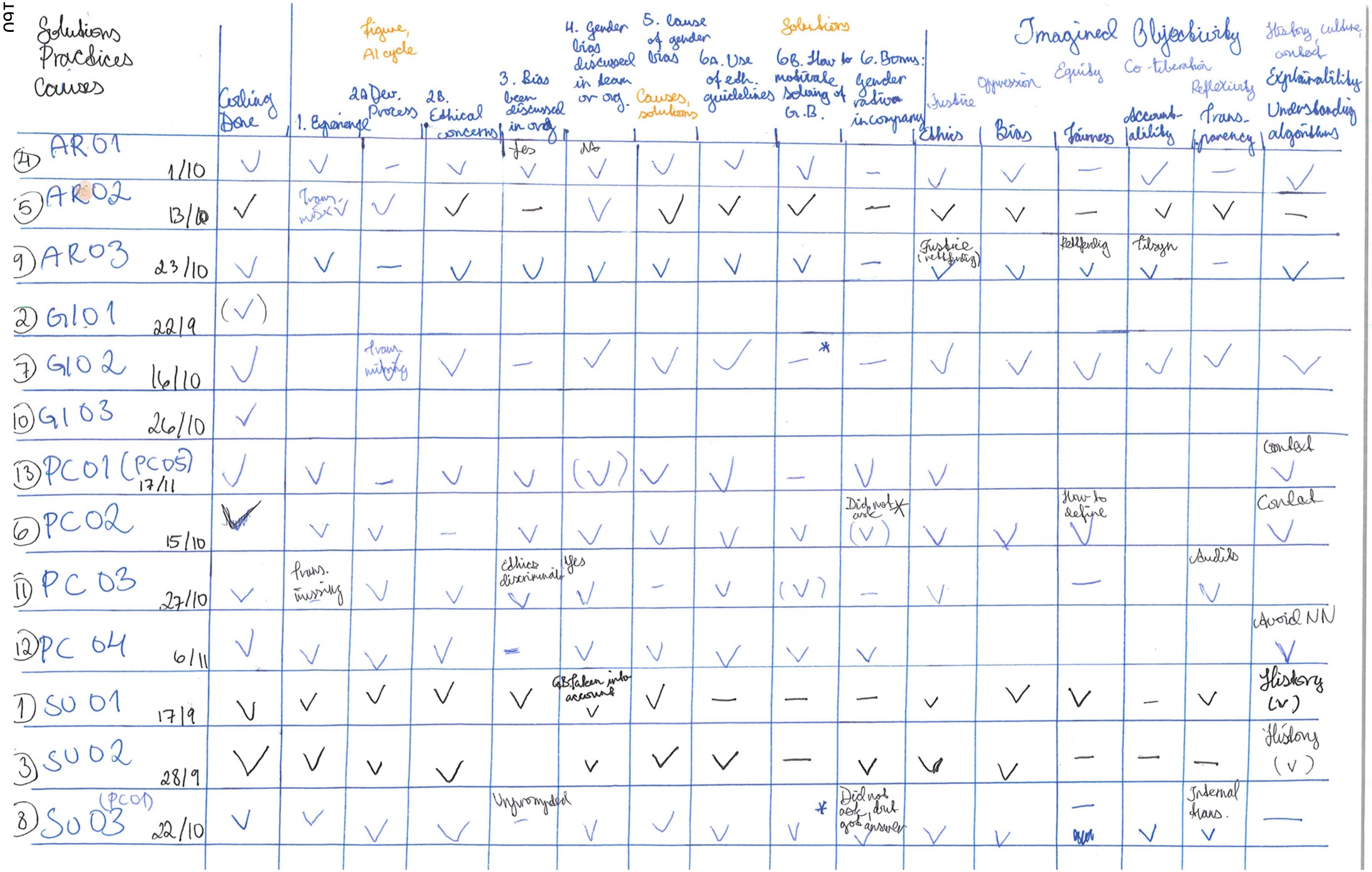
□ I give consent for my personal data to be processed until the end of the project date.
□ I agree that research data gathered for the study may be published provided my name or other directly identifiable information is not used, unless other consent have been given.
□ I give consent for sensitive personal data to be collected and archived after it has been anonymized.

(Name) (Signature) (Date)

Appendix F - Coding Table on Paper



		Instrumentalism |- - -	-	-	- CTOT -	- - -	| Leverage Points		Gendertheory	Standpoint theory, Strong objectivity			
	Coding done	Tech is neutral	Sep. of tech vs. values/ethics	Who has Power?	Tech. heritage	Strong vs. weak designers	Optimize for?	Motivation goal?		Language bias	View on objectivity	Equity	Intersectionality	
AR01	✓						✓	Curiosity ✓		–		–	–	
AR02	✓													
AR03	✓		CTOT?	✓	Tech solutionism ✓	Strong ✓	Better world ✓	Curiosity, better world ✓		✓	Object. outreach ✓	How to correct for bias ✓	=	
GI01	(✓)													
GI02	✓													
GI03	✓													
PC01 (PC/05)	✓												White male data ✓	
PC02	✓			They do surely ✓		Strong ✓		Profit + Human Rights ✓						
PC03	✓			Market ✓		Weak ✓	Customer experience ✓	Customer ✓		–				
PC04	✓	Yes ✓		Big tech ✓		Weak? ✓								
SU01	✓	✓	✓	Corp. ✓	Tests ✓	Strong ✓	Performance ✓	Better world ✓		✓	✓			
SU02	✓	✓	✓	✓	Techfocus ✓	Strong ✓	Surviving, ~~Idealism~~ ✓	Better world ✓		✓	✓		✓	
SU03	✓	Value-laden ✓	Some AI's more prone to bias ✓	Citation system, conf. ✓		Strong ✓	Surviving ✓	Better world ✓		Aware ✓	Science never objective ✓	–	Diversity in race and gender ✓	



		Figure, AI cycle					Causes, solutions	Solutions			Imagined Objectivity Justice	Oppression	Equity	Co-liberation	Reflexivity	History, culture, context Explainability
Solutions Practices Causes	Coding Done	1. Experience	2A Dev. Process	2B. Ethical concerns	3. Bias been discussed in org	4. Gender bias discussed in team or org.	5. Cause of gender bias	6A. Use of eth. guidelines	6B. How to motivate solving of G.B.	6. Bonus: Gender ratio in company	Ethics	Bias	Fairness	Accountability	Transparency	Understanding algorithms
(4) AR01 1/10	✓	✓	—	✓	Yes ✓	No ✓	✓	✓	✓	—	✓	✓	—	✓	—	✓
(5) AR02 13/10	✓	Trans. miss x ✓	✓	✓	—	✓	✓	✓	✓	—	✓	✓	—	✓	✓	—
(9) AR03 23/10	✓	✓	—	✓	✓	✓	✓	✓	✓	—	Justice (rettferdig) ✓	✓	Rettferdig ✓	Tilsyn ✓	—	✓
(2) GI01 22/9	(✓)															
(7) GI02 16/10	✓		Trans. missing	✓	—	✓	✓	✓	— *	—	✓	✓	✓	✓	✓	✓
(10) GI03 26/10	✓															
(13) PC01 (PC05) 17/11	✓	✓	—	✓	✓	(✓)	✓	✓	—	✓	✓					Context ✓
(6) PC02 15/10	✓	✓	✓	—	✓	✓	✓	✓	✓	Did not ask ✗ (✓)	✓	✓	How to define ✓			Context ✓
(11) PC03 27/10	✓	Trans. missing	✓	✓	Ethics discriminate ✓	Yes ✓	—	✓	(✓)	—	✓		—		Audits ✓	
(12) PC04 6/11	✓	✓	✓	✓	=	✓	✓	✓	✓	✓						Avoid NN ✓
(1) SU01 17/9	✓	✓	✓	✓	✓	GB taken into account ✓	✓	—	—	—	✓	✓	✓	—	✓	History (✓)
(3) SU02 28/9	✓	✓	✓	✓		✓	✓	✓	—	✓	✓	✓	—	—	—	History (✓)
(8) SU03 (PC01) 22/10	✓	✓	✓	✓	Unprompted —	✓	✓	✓	✓ *	Did not ask, but got answer ✓	✓	✓	—	✓	Internal trans. ✓	—