Capturing Power:

Feminist Considerations about Machine Learning Fairness

Alexandru-Nicolae Postu[1]
Supervisor(s): Sarah Carter[1]**, Jie Yang**[1]**, Stefan Buijsman**[2]
[1]EEMCS, Delft University of Technology, The Netherlands
[2]TPM, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Alexandru-Nicolae Postu
Final project course: CSE3000 Research Project
Thesis committee: Jie Yang, Sarah Carter, Marcus Specht, Stefan Buijsman

Abstract

Machine Learning (ML) algorithms have the potential to reproduce biases that already exist in society, a fact that leads to scholarly work trying to quantify algorithmic discrimination through fairness metrics. Although there are now a plethora of metrics, some of them are even contradictory, making fairness become a problem of knowing which measurement to choose over another. Consequently, scholars began considering that fairness should be discussed by placing algorithms in their social contexts. Since (1) these social aspects are related to structures of discrimination and (2) feminism aims to criticise discrimination against the marginalised, I introduce the possibility of analysing the social context of ML algorithms through a feminist lens. By doing this, I highlight social and political aspects that are equally important to consider for a faithful discussion on fairness: corporate lobbying, the lack of diverse hiring which leads to fairness discussions that do not consider the experiences of marginalised groups and, lastly, the broader context that an algorithm is used in. Moreover, I emphasise how feminist ethics of care constitute an essential framework for a conversation about actually implementable fairness solutions, since it shows the need to listen to both the marginalised community and to the developers who might want to build fairer ML but currently cannot. Having built a bridge between the hegemony and the feminist camp, I highlight how Northpointe's (now Equivant's) Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm can be considered biased against black people. Through this, I illustrate how feminist considerations bring clarity to fairness debates by helping choose a fairness metric or by claiming that an algorithm is unfair by nature and should be abolished. To follow, I use the same feminist critiques to draw attention to the possible weak points of current sociotechnical solutions. For instance, the EU AI Act risks being too susceptible to company lobbying, leading to not strict enough regulations. Furthermore, the AI committee should ensure that they hire a diverse group of people in order to develop regulations that positively consider all marginalised groups. Lastly, I highlight how ethics education is essential for creating a new generation of responsible engineers. Considering this, I emphasise the urgency of making ethics courses at TU Delft (and not only) more interdisciplinary by interacting more with critique points coming from the social sciences. This will open up possibilities for more research tackling fairness from a multitude of perspectives.

1 Introduction

With the improvement of Machine Learning (ML) technology, both private and public institutions have been employing it to speed up and remove bias from processes. These algorithms now decide whether to hire applicants or not (Gonzalez et al., 2022), predict if patients might have kidney stones (Caglayan et al., 2022) and even predict the crime recidivism possibility of a person (Dakalbab et al., 2022). However, human bias made its way into algorithmic bias. The same algorithms exhibited preferences for male applicants (Dastin, 2018), were more likely to incorrectly classify black people as not having kidney stones (Vyas et al., 2020) and even predicted them as more likely to commit crimes again (Angwin et al., 2016). The findings about these biases caused a spike in the amount of research that investigates fairness and discrimination in the field of ML, with over 1000 papers per year being published since 2021 (Kheya et al., 2024).

This led to scholars mathematically defining the concept of fairness for ML algorithms. However, there is currently no agreed definition and some fairness metrics are even contradictory (Richardson and Gilbert, 2021, Mehrabi et al., 2019, Barocas et al., 2023). This body of work, which focuses on technical solutions and metrics, mainly comes from a Computer Science background and makes up the knowledge that is predominantly listened to. Consequently, I will refer to this produced knowledge as ‘hegemonic knowledge‘ [1]. Recently, there has been literature which recommends that fairness ought to be defined depending on the domain the algorithms are applied on: in other words, contextually (Corbett-Davies and Goel, 2018, Green and Hu, 2018). Expanding on this context, scientific work which asks for sociotechnical solutions instead of purely technical ones has recently emerged (see Green and Hu, 2018, Richardson and Gilbert, 2021, Mehrabi et al., 2019). In other words, the hegemony itself seems to change from purely technical to sociotechnical fairness solutions.

However, the field of ML fairness relates specifically to the fact that algorithms discriminate marginalised communities. This category of people is the specific focus of feminist philosophy, which evaluates how mechanisms of society lead to uneven distributions of power and resources. Thus, there is value in analysing ML systems from such a feminist perspective. Applied to ML, such literature focuses on the context that algorithms are employed in and on their potential to perpetuate and even accentuate inequality. The feminist focus is put less on the technical design of the algorithms and more on questioning whether the design of specific ML algorithms is ethical to begin with (see, for instance, McInerney, 2023).

In this paper, I aim to investigate the value that bridging the hegemony and the feminist camps brings when discussing fairness in ML and ways of achieving it. To do so, I will firstly provide a description of the hegemonic fairness literature in Section 2. In particular, I summarize the predominant existing fairness metrics and some common critique points. In Section 3, I will apply feminist epistemology and black feminist rhetoric to the concept of ML fairness, highlighting how this puts the unethical practices of Silicon Valley corporations, the need for more diverse hiring practices and the idea that some algorithms are unethical by default at the forefront. To highlight the value a feminist approach adds to discussions about ML fairness, I will dedicate Section 4 towards building a bridge between the hegemonic and the feminist frameworks by using the notion of ethics of care. To highlight the value of feminist critique, I will first highlight the additional considerations that they bring when discussing the COMPAS debate, in Section 5. This, I argue, could be applied to similar investigations into whether a system is fair or not. To follow, in Section 6 I use feminist critique to imagine possible long-term solutions towards improving fairness in the ML field. To do so, I use the EU AI Act as a starting point, highlighting extra considerations that a feminist approach will have. Lastly, I conclude in Section 7 that analysing the social context of an algorithm helps inform what is fair, that the EU AI Act needs to be wary of lobbying and that interdisciplinary ethics education is necessary to push the hegemony towards more inclusive fairness.

2 Hegemonic Perspective

Through this section, I aim to summarize the manner in which fairness is interpreted by the hegemony. To do so, I identify fairness metrics that have been highlighted by more than one literature survey. Then, I identify common points coming from within the hegemony.

[1] Through ‘hegemony‘, I convey that there is a dominant shared system of ideas and ethics within a period of time (Lauderdale & Amster, 2008). This predominant set of ideas is then used to preserve the status quo, making it difficult for opposing views to emerge (Bates, 1975). In the context of this paper, I use the term to illustrate that the fairness ideas in Computer Science academia potentially drown out critique points coming from other academic disciplines. Furthermore, I highlight how large corporations profit from this hegemony.

2.1 Fairness metrics

The literature contains a significant amount of fairness metric definitions and attempts to categorise them. In their paper, Richardson and Gilbert survey and aggregate a significant amount of fairness metrics (2021). Namely, they mention that Mehrabi et al. identify 10 different types of fairness (2019), while Barocas et al. find 19 (2023) and Verma and Rubin, 20 (2017). Similarly, Ferrara enumerates 5 different fairness metrics categories (2023). In addition, Kheya et al. highlight 17 fairness metrics that are then split in 7 categories (2024). Lastly, Ruf and Detyniecki identify 10 different definitions (2021), with 3 categories. By only considering metrics that are present in at least two of these papers, I myself gather 17 fairness metrics, which can be found under Appendix A. Ideally, a developer should be able to evaluate how fair their algorithm is by picking the fitting definitions for the problem and by then fixing biases that emerge from the statistics. However, through this brief surveying I already highlight that different people may not find the same metrics, as different surveys with the same goal have not been able to, either. This, I concur, hints at a current lack of consensus on what metrics are actually relevant.

For the scope of this paper, I highlight two metrics that will be important in Section 5.

- **Predictive equality** (Kheya et al., 2024; Ruf and Detyniecki, 2021; Barocas et al., 2023)
  According to Kheya et al., the false positive rate is equal across groups, regardless of sensitive attributes. In other words, this metric measures the probability of somebody being mistakenly classified as a member of the positive class (2024).

- **Predictive parity** (Kheya et al., 2024; Verma and Rubin, 2017; Ruf and Detyniecki, 2021; Barocas et al., 2023)
  According to Kheya et al., the positive predictive value across groups is equal, regardless of sensitive attributes. Simply, this metric represents the probability that a person is correctly classified as a member of the positive class (2024).

Most importantly, Chouldechova shows that predictive equality (introduced as error rate balance) and predictive parity cannot be both satisfied at the same time (2017), meaning that one must be chosen over the other.

2.2 Critique points

The contribution of this paper stems from intersecting the hegemonic and feminist perspectives, highlighting the possibilities of this common ground. Thus, I only identify the critique points which interact with the notions of context and sociotechnicality, making them fitting starting ground for an intersection with the social feminist critiques that will follow later.

Firstly, there are too many fairness metrics (Mehrabi et al., 2019; Richardson and Gilbert, 2021) which cannot be satisfied all at once, requiring trade-offs (Richardson and Gilbert, 2021; Majumder et al., 2023; Ferrara, 2023). Even so, cases of ML discrimination that are not covered by any of these metrics still exist (Holstein et al., 2018). This abundance of metrics makes it difficult for developers to choose what they need for their application (Mehrabi et al., 2019). To further complicate this problem, different stakeholders interpret fairness differently (Ferrara, 2023). Therefore, it seems impossible to achieve a single mathematical fairness solution that fits the needs of all existing stakeholders to begin with (Ferrara, 2023; Ruf and Detyniecki, 2021).

Secondly, the technical aspects of fairness are emphasised to the detriment of the social ones. According to Green and Hu, this reduction from the sociotechnical to the purely technical leads to a misinterpretation of the results (2018). They further argue that this misinterpretation leads scholars to see the conflicting fairness metrics as proof that complete fairness is unachievable. However, this simply reconfirms that the problem space and the solution have a social component attached to them. Corbett-Davies and Goel also highlight the limitations of dominant mathematical solutions by acknowledging that statistical evaluations do not imply an increase in well-being for the minority groups that are affected (2018).

Thirdly, fairness as a concept is understood too abstractly, leading to analysis that does not meaningfully consider context. Selbst et al. introduce the concept of "formalism trap" (p. 61) [2], while highlighting that the choice of fairness metrics should be dictated by socially-informed choices (2019). John-Mathews et al. further argue that the hegemony tries solving the fairness problem by using the given dataset but not considering that the whole context that the ML solution is deployed in, dataset included, is rooted in inequality (2022). To exemplify this, the authors mention that a team of white ML developers trying to fix a system biased against black people might improperly select the fairness metrics due to their own biases.

Based on these critiques, I argue that fairness cannot be achieved through statistical metrics alone and that context is essential. Without social context, metrics fail to account for the well-being of the marginalised communities.

3 Feminist Perspective

Feminist work can be understood as a critique of hierarchical gender structures and of their interplay with race, class and heterosexism dynamics (Allen, 2022). In other words, feminist work generally criticises aspects of society with the aim of exploring new solutions to combat discrimination. In this section, I elaborate on the need for a feminist approach, so that I can then employ branches of feminist philosophy to elaborate on three crucial aspects to contextualizing if an algorithm is fair. These accounts are namely: who are the algorithms developed for, by whom and for what.

3.1 Why feminist philosophy?

Firstly, feminist philosophy has historically interacted with the hegemony and turned out to be correct over time. For instance, Bender et al.'s paper cited several concerns about Neural Language Processing (NLP) technologies, such as negative environmental impact, financial costs and bias-related problems that arise when the dataset used gets too large (2021). This paper was written before the deployment of ChatGPT, one of the currently most talked about NLPs. Nonetheless, Gebru's NLPs concerns proved correct for this tool. For instance, ChatGPT is indeed not environmentally friendly, as it uses significant amount of water (George et al., 2023). In addition, research shows that ChatGPT might express positive biases for left-leaning viewpoints, in spite of it claiming to hold no political opinions at all (Rozado, 2023).

[2] "Failure to account for the full meaning of social concepts such as fairness, which can be procedural, contextual, and contestable, and cannot be resolved through mathematical formalisms." (Selbst et al., 2019, p. 61)

Secondly, there is merit in interacting with political activism and philosophical literature even if the proposed solutions might sound far-fetched, as these discussions spark new remarks about the currently existing social structures. For instance, abolitionist feminism critiques the existence of predictive policing algorithms, arguing that the idea of such systems should be abandoned altogether (Noble, 2020). Although this idea comes from theories controversially claiming that prisons should be abolished, some predictive policing algorithms themselves are now labeled unacceptable by the EU AI Act (Yakinova & Ojamo, 2024), showing how there was value in this claim.

Lastly, feminist philosophy has a better grasp on issues of social inequality. For ML algorithms, this inequality can manifest itself twofold. First, it can appear in the social environment that the algorithm is developed in. In such an environment, biases can be introduced through having non-diverse groups of developers or through organisations which perpetuate discriminatory practices. Second, the algorithm itself produces social inequality, as its results can potentially lead to denying people of resources that they should have had access to. In spite of the decision being made technically, the impacts are social. For instance, not being hired for being a woman has the same implications for the person rejected, regardless of whether it was a person or an algorithm that made the decision.

3.2 Feminist critique points

In this section, I will highlight a variety of feminist work, consequently highlighting three aspects that must be considered when discussing fair ML solutions. In particular, I develop a thought framework which asks for who, by whom and for what an algorithm is developed. These three components will help provide insight about currently existing cases of ML unfairness and about sociotechnical ML fairness solutions.

3.2.1 Who are the algorithms developed for?

Through this argument, I wish to highlight that companies prioritize expansion of capital over ethics and the environment. Thus, we should pick the fairness metrics and solutions that advantage the marginalised communities to the detriment of corporations. This tackles the problem that, for the same algorithm but different stakeholders, the desired fairness metrics can be different (Ferrara, 2023, Ruf and Detyniecki, 2021) by introducing a prioritisation of stakeholders. Indeed, favouring the marginalised does not satisfy all stakeholders. However, it does favour the stakeholder that the concept of fairness should be centered on, also according to Rawl's difference principle (Waldron, 1993) [3].

To argue for this, I use black feminist theory. This framework helps us think about the quality and content of social hierarchies, while also accounting for race as a consideration, an aspect that is normally neglected (Noble, 2020). Through a black feminist lens, one also ensures that no marginalised group is left behind, since by liberating "(trans) Black women from algorithmic oppression, we liberate everybody from algorithmic oppression" (Hampton, p. 3). Furthermore, this framework allows for the understanding that the current decision-making protocols favour the large corporations, perpetuating global and social economic inequality (Noble, 2020). Lastly, This point is considerably important considering that companies such as Microsoft, Amazon and Google recorded increases of $211 Billion (Nadella, 2023), $61 Billion (Jassy, 2024) and $25 Billion (Sundar, 2024) respectively.

[3] This principle is based on the idea that inequality should be permitted in society if and only if it is meant to favour the most marginalised (Waldron, 1993).

Hampton argues that power imbalances are deeply embedded in societal systems and institutions and this includes Silicon Valley companies (2021). In addition, companies are driven to 'act ethically' solely by the possibility of gaining capital. For instance, corporate gender equality is majorly influenced by the belief that hiring more women is simply more profitable for the company (Roberts, 2014), instead of the belief that it is morally just to do so. Thus, I argue that companies prioritise capital gain over ethics. This is deeply problematic: if fairness is open to so many metrics, interpretations and varying risks, could it not be that Silicon Valley companies will simply pursue the most economically profitable ML fairness solutions to the detriment of marginalised people?

Numerous cases point out that this is true. For instance, Amazon pushed for facial recognition tools in spite of two black women's attempt to showcase the harmful biases that these exhibited (Singer, 2019). Equivant„ refuted the results of independent investigations which worried that their algorithm, COMPAS, is more likely to predict black people as likely to commit crimes again (Dieterich et al., 2016). More generally, companies employ lobbying practices that can shape the hegemony's research of ML ethics according to their interests (Weinberg, 2022; Phan et al., 2021). For instance, 'Partnership on AI to Benefit People and Society', a group founded by Microsoft, Google, Facebook, IBM and Amazon in 2016, insisted that risk assessment can be regarded as a technical issue. This is in spite of efforts of other grassroots movements, which are comprised of people from marginalised backgrounds who are directly impacted by risk assessment tools, to argue against risk assessment AI technology (Phan et al., 2021). On the same note, Saltelli et al. investigate cases of corporations retorting to regulatory capture in order to strategically influence aspects of science (2022)[4].

Through this argument, I emphasize the need to consider that multinational companies act inconsistently when it comes to ethical practices. The cases highlighted above show a contradiction between the values that Silicon Valley corporations claim publicly and the actions that they commit once not faced with social resistance. For the scope of this paper, the idea is that corporate intentions should be doubted when it comes to ethics.

3.2.2 Who are the algorithms developed by?

The goal of this second argument is to highlight that more diverse hiring is a necessary, but not effective on its own, step towards ML fairness solutions that support marginalised groups. The hegemony is now predominantly composed of cis, heterosexual, white, able-bodied men (Gebru, 2020a, West et al., 2019), fact which can influence the research output with regards to what is fair and not. This is particularly problematic given the general, but false perception over research as abstract and objective (Tuck & Yang, 2014).

According to feminist epistemology, the atomistic [5] epistemological model represents an obstacle to the production of knowledge which can accurately illustrate systematic power relations (Grasswick, 2018). In other words, assuming one single possible way of producing knowledge fails to take into account power hierarchies. In particular, the concept of partial

[4] These cases exhibit one commonality: companies use their scientific authority, internal hierarchies and resources to steer scientific production in a direction that fits their business plans. Preserving the hegemonic knowledge is intrinsic to holding this practice of lobbying disguised under ethical research. Thus, it is crucial to offer more platform to non-hegemonic forms of knowledge such as feminist philosophies, as they can be more critical of such practices. In order to ensure that ML regulations respect the needs and integrity of the marginalised, it is crucial to watch the currently developing EU AI committee critically.

[5] Through an atomistic model, it is understood that a person's social location does not matter in the production of knowledge (Grasswick, 2018), fact which effectively negates knowledge based on lived experience. I tie this idea to the hegemony, as both concepts inherently place the currently produced knowledge into a position of dominance that excludes the marginalised.

objectivity (Grasswick, 2018) emphasises that knowledge produced by including people from minority backgrounds is superior to the hegemonic one, as it encompasses more lived experiences. This larger number of lived experiences leads to less distorted knowledge production.

Following this idea, I argue that a feminist perspective on fairness in AI algorithms is crucial, as the ones discriminated by AI come from marginalised communities: they are black, women, queer, of immigrant background and so on. If one ought to attempt to achieve fairness, then the people from the discriminated categories should have a say in the production of fairness literature. That way, the research output will have more focus and urgency on not discriminating marginalized communities. However, the AI field is right now predominantly made up of white, cisgender male developers (Gebru, 2020a, West et al., 2019) - in other words, of the people who are not negatively impacted by unfair AI to begin with. Indeed, the hegemonic knowledge can be considered distorted, since it fails to encompass the lived experiences of the disproportionately affected minority groups.

Finally, one can understand the conflicting fairness metrics as proof that knowledge is situated. Given that there is no fairness metric or solution that appropriately fits all stakeholders, the choice of fairness metric is influenced by the authors' social location. In other words, the conflicting opinions over the right fairness metrics can be understood as a manifestation of stakeholders belonging to different social locations. Given the predominance of white, cisgender and heterosexual men in the AI industry, both within research and corporations (West et al., 2019, Bender et al., 2021), it can be argued that the current research about fairness is predominantly influenced by the perspective of the people highest located in the power hierarchy. The first and smallest step towards reaching fair ML, thus, is to start actually including the discriminated people in the development process.

I will close this argument by emphasizing that diverse hiring practices will not solve anything if applied as a standalone solution. Through a black feminist lens, corporations are understood as relying on hierarchical distributions of power. On this note, Acker claims that organizations use their power to ensure employees act in compliance with the company's goal. Hampton goes further, arguing that companies can use their internal power hierarchies to shift blame from the upper management to individual employees who are already marginalised, making more diverse hiring an incomplete solution (2021). Not coincidentally, employees do not advocate for AI fairness in fear of negative consequences to their careers (Madaio et al., 2020). Timnit Gebru's lay-off from Google validates this claim. Following Google's refusal to publish a paper critiquing large-scale NLPs (Gebru, 2020a), Gebru asked for clarifications for the rejection, threatening to resign otherwise. Afterwards, she allegedly got suspended from her employee account without any notice (Gebru, 2020b).

3.2.3 What are the algorithms developed for?

The third and last significant consideration to make is the purpose of ML algorithms. Critically evaluating the purpose can even lead to scraping the developing of algorithms altogether, as the pure existence of some can be deemed as unfair depending on the social theory employed to make this analysis.

As Weinberg points out, current ML fairness research is produced within the same social, cultural and economic context that has caused the existing uneven distributions of power to begin with (2022). This is important to remember on two accounts. Firstly, a social system riddled by discriminatory practices will not be fixed by a ML algorithm. People's biases are informed by stereotypes which do not accurately reflect reality. The same, however, holds for ML algorithms: they are trained on datasets which are biased as a consequence of past discriminatory practices (Draude et al., 2019), leading to systems with the same biases exhibited by a less transparent system (von Eschenbach, 2021). Secondly, ML/AI solutions risk accentuating the currently existing hierarchical, uneven power dynamics in society (Weinberg, 2022). This conclusion has also been, in fact, reached by the hegemony. Zou and Khern-am-nuai conclude that algorithms deciding the outcome of mortgage applications actually accentuate the racial bias pre-existing in the dataset (2022).

This consideration is important to make when discussing if a system is fair or not. If the context that a ML system is deployed in already finds itself under ethical scrutiny, this context should first be re-evaluated and changed before implementing a system with less transparency and more difficult to implement accountability measures.

4 Bridging the Perspectives

When applied to the field of ML, both hegemonic and feminist perspectives strive to achieve fairness. These different perspectives exhibit common ideas that, once written down, can both improve existing solutions or generate new ones. Firstly, I will argue why this intersection can be fruitful. Secondly, I will elaborate on how this intersection of perspectives can be used.

4.1 Why build the bridge?

On one hand, the hegemonic literature helps provide implementation solutions towards building fair ML algorithms. Developing a ML algorithm requires technical expertise, as the ML system itself is technical. For instance, a critique point of deep learning systems is that the decision making process is too opaque to the people impacted (von Eschenbach, 2021). A solution to the lack of transparency in the decision making processes of ML algorithms is explainable AI (XAI), a field of research to which significantly more attention has been paid to since 2017 (Xu et al., 2019). For developing XAI solutions, technical expertise is needed, as one needs to understand the algorithm before being able to translate the output results into more understandable language.

Moreover, there are feminist frameworks which aim for reshaping societal structures by attentively listening to all involved stakeholders. In particular, the philosophy of care, as described by Fisher and Tronto, represents acts performed to "maintain, continue and repair our bodies, our selves and our environment" (Fisher et al., 1990, p. 40). In particular, I remind of the concept of "caring about" (Tronto, 1998, p. 16), which emphasizes a need to genuinely listen to all people of our society, in order to understand both their articulated and unarticulated needs. This philosophy is particularly beneficial for bridge building, as it emphasises a need to understand the difficulties that everyone is facing. Applied to the context of this paper, it is very important to draw a distinction between Sillicon Valley corporations and the Sillicon Valley corporation workers. The former is a system in need of reform, whereas the latter represents people who can help towards achieving this reform. Following the ethics of care, intersecting the hegemony and the feminist camp means more sociotechnical solutions and regulations that are also implementable by a ML developer. Experts have been surveyed to find fairness difficult to implement in their ML systems (Holstein et al., 2018), meaning that more work must be put into different, more applicable fairness solutions. Coming up with such solutions will require listening to the experts that implement the ML systems. However, this needs to be done in a manner that can be isolated from the corporate pressure that Madaio et al. (2020) highlighted, so as to ensure honest answers. Nonetheless, it is crucial that interacting with the experts does not lead to disregarding the needs of the marginalised again.

Finally, bridging this gap means that researchers can come up with sociotechnical solutions of higher quality. Individually, both the technical and the social foundations have strong backgrounds, but diverging outcomes. Once combined, these two camps can influence each other and converge towards new possibilities. Take the XAI example from above: there is an identified need for transparency when it comes to decision-making processes, as identified by social sciences. However, providing such solutions requires technical expertise that belongs to the hegemony.

4.2 How can the bridge be built?

The intersection of perspectives lies in the belief that fairness should be understood contextually. John-Mathews et al. highlight criticism on how solutions and metrics do not consider context nearly enough, with some fairness solutions actually accentuating unfairness if applied mistakenly (2022). I argue that a feminist approach to ML directly fills this context gap. To build the bridge, I provide more analysis on the context of an algorithm, by considering for whom, by whom and for what a ML solution has been developed.

This context brings two novel points. First, it helps bring additional clarity on debates about algorithmic discrimination, because feminist philosophy has a better grasp on the long-lasting impact and consequences of discriminatory decisions. In this sense, discussions about how to quantify the fairness of an algorithm could be better informed by feminist thinking, as this focuses more on the social and structural power dynamics that the algorithm can reinforce. Moreover, such an approach might even lead to concluding that a ML system is unfair by definition. I highlight how this would actually work in Section 5. Second, feminist critique can help highlight additional weaknesses in new fairness solutions, especially if these are sociotechnical. Highlighting current weaknesses of ongoing solutions can spark discussons which lead to improvement. I dedicate Section 6 to this process, opening up the possibility for further research questions.

5 COMPAS: Case Study

In this section, I aim to highlight how contextualising an algorithm from a feminist lens aids in discussing fairness in a way that supports the discriminated. To do so, I follow the COMPAS debate between Equivant and ProPublica and draw two different conclusions. Using the argument from 3.2.2, I argue that the fairness metric chosen should be Predictive Equality. By considering the purpose of the algorithm as explained in 3.2.3, I further explore the possibility that COMPAS as an algorithm might be unfair de facto.

5.1 COMPAS Debate

The COMPAS algorithm designed by Equivant predicts the recidivism rate of previously convicted people. This algorithm, which played a role in decision-making processes in states such as Florida, is both inaccurate and discriminatory according to ProPublica's investigation. COMPAS was twice as likely to label black people as having a high risk of recidivism when compared to white people (Angwin et al., 2016). Following this investigation, Equivant refuted the claims, arguing that the fairness metric chosen by ProPublica was incorrect for this use case (Dieterich et al., 2016). This circles back to a problem existing in the hegemonic research: the contradicting metrics. Following a purely statistical analysis, a contradiction had already been reached: there are two correct formulas which do not produce the same result on the exact same problem. However, socially understanding what these metrics mean for the impacted people can lead to a better choice of metric.

5.2 Incompatible Metrics

Propublica's claim is that COMPAS is unfair, since the false positive rate is twice as high in the case of black people (Angwin et al., 2016). This metric is introduced in academia as predictive equality. However, Dieterich et al. argue that it is not predictive equality which should be achieved, but predictive parity. Since COMPAS satisfies predictive parity, it is allegedly not biased towards black people (2016). I have introduced both of these metrics in Section 2.1.

Contextualized, predictive parity means that COMPAS correctly predicts someone as likely to commit a crime again in just as many cases, regardless of race. Although Equivant dismisses predictive equality as irrelevant (Dieterich et al., 2016) [6], not satisfying this metric does, indeed, mean that COMPAS mistakenly predicts a black person as likely to commit a crime again twice as often (Angwin et al., 2016). This is particularly problematic considering that the algorithm is not transparent, making it difficult to reason about its decisions (Rudin et al., 2020). I argue that it is unfair for an algorithm to be twice as likely to mistakenly punish somebody on the basis of being black, especially given the prevalence of racism in the US prison system (Davis, 2000; Hampton, 2021; Demers, 2013).

5.3 For whom and what is the fairness metric chosen?

By considering who the algorithm is made for, I argue that the choice of fairness metric is influenced by the stakeholder that each side wishes to represent. First, Equivant is a company specializing in designing 'software for justice', providing assistance to courts, attorneys, supervisions and custodies (Equivant, 2020). Thus, their interest is to gain profit through collaborating with the USA juridical system. Second, ProPublica is a non-profit investigative platform, aiming to serve public interest (ProPublica, 2017). Thus, they represent the defendants, who are at the receiving end of the mistaken decisions. Not satisfying predictive equality does, on this note, suggest that black people at twice the risk of being wrongfully convicted. From a feminist perspective, the stakeholder that should be most accounted for are people of minority status: in this case, black people. Finally, I argue that we should choose predictive equality (a metric representing the interest of racial minorities) over predictive parity, which represents corporate interests.

An interesting conclusion can be drawn by considering the purpose of COMPAS. By using the abolitionist feminist idea that asks for the dismantling of predictive policing algorithms altogether (McInerney, 2023), one can argue that COMPAS is unfair regardless of metric and should be abolished. This ML system is used to influence decisions related to convicting people in prisons or not. However, US prisons are understood by anti-carceral feminists such as Carlton as institutions which perpetuate the same oppressive violence that they claim to combat by relying on oppressive attitudes based on gender, sexuality, race and class (2016). In addition, Bland argues that such predictive policing algorithms do not even fix the existing

[6] In particular, they claim that predictive equality is not correct because it does not account for different base rates of recidivism (Dieterich et al., 2016). However, these base rates could be understood as a larger byproduct of systemic inequality.

issues to begin with, representing nothing more than a solution swap (2020). Going even further, some scholars argue that predictive policing algorithms create problems that did not exist in the past. For instance, McInerney argues that predictive policing alorithms send the impression that crimes such gender-based violence are a naturally occurring phenomenon that always exist (2023). On a similar note, Salman argues that COMPAS adds a fake layer of objectivity to inherently racist decision-making (2024). These two problems are particularly problematic, as ML algorithms add the additional problem of behaving akin to a black-box, which will not be understood by the people who have to face its decisions (Bland, 2020).

Summing these critiques up, I argue that COMPAS risks having a role in accentuating systemic discrimination. This ML system non-transparently makes predictions which the impacted people can experience as racist. It does so under the pretense of 'mathematical objectivity', making it difficult for the underprivileged to defend themselves. Lastly, it does this all while being part of a system based on discriminatory practices.

6 Solutions

In this section, I look at sociotechnical solutions that are currently being explored. By posing the questions from 3, I identify aspects that need to be considered when further developing these solutions. Concretely, I critically analyse the European Union (EU) AI act in 6.1 and the evolution of ML education at TU Delft in 6.2.

6.1 The European Union AI Act

The European Union (EU) AI act is the first attempt at large-scale regulation of AI systems. It was first proposed in April 2021 (Veale & Borgesius, 2021) and then formally adopted in May 2024 (Mammonas, 2024). The AI act classifies AI systems in terms of risk and labels them as unacceptable, high-risk, limited-risk and minimal-risk. This classification influences the number of restrictions that an AI developer or deployer must abide to. In particular, unacceptable systems are prohibited on EU territory, whereas high-risk systems need to abide to a strict set of rules. Minimal-risk systems remain largely unaffected (Mammonas, 2024). Section 3 highlights the need for external intervention into companies practices, particularly by following considerations such as for whom (see 3.2.1) and for what (see 3.2.3) a ML algorithm may be designed. In that sense, the EU AI Act is a step in the right direction.

First, all ML solutions exemplified so far are created within a corporate context. Considering that Silicon Valley corporations are more likely to prioritise profit over in-depth ethical considerations when making choices about the algorithms they design, there is no guarantee that leaving companies to manage their own fairness metrics and solutions will lead to a change in the status quo (John-Mathews et al., 2022). Second, critically evaluating the usage contexts of the ML algorithms could lead to considering some ML solutions unfair from the get-go. For instance, suppose that general and academic opinion converges towards thinking that developing predictive algorithms for juridical systems is unethical. For companies whose business models rely on ethical algorithms, thus, such conclusions mean that their business should not exist to begin with. Thus, I conclude external regulation is a step in the correct direction. Nonetheless, the AI act has been criticised as not effective enough in safeguarding society against unethical ML systems. I agree with this on three accounts.

For one, stricter restrictions of the draft have been removed, allegedly due to corporate lobbying practices. This is claimed by Veale and Borgesius, who says that the 'High-Level Expert Group on AI', meant to advise on the EU's AI strategy, was confronted with industry lobbying (2021). In particular, the AI draft leaked in 2018 required the providers of AI high-risk systems to ensure that overseers would be allowed to scrap the system at their own discretion, without facing any consequences (Veale & Borgesius, 2021). Considering who algorithms are developed for is important in two ways. On one hand, it allows me to conclude that choosing to remove this measure is regrettable, as it would have empowered employees to act ethically, even if in opposition to the company's management. This could have helped combat the fact that employees do not feel safe advancing for ML fairness measures and practices in their own companies (Madaio et al., 2020). Furthermore, choosing to remove a tool would have given agency to non-management illustrates the companies clear interest in preserving a hierarchical, power-based structure. On the other hand, the consideration allows me to support the lobbying claim. As highlighted previously (3.2.1), large-scale companies are indeed lobbying for their desired outcome in fairness research. Combining this with the fact that companies historically have lobbied for favourble political outcomes (Bernhagen & Bräuninger, 2005), I argue that lobbying can interfere with the AI act, too.

Secondly, I question who defines the risk-level of a system to begin with. Mahler sees merit in using risk analysis as a methodology for regulation, but they argue that it is not clear how the risks of technologies were assessed (2022). Most importantly, the law-making processes and risk acceptance can be influenced by political, economical, social and other considerations (Mahler, 2022, Fraser and Bello y Villarino, 2023). I argue that company lobbying could be placed among those considerations. Though not explicitly mentioning lobbying, Fraser and Bello y Villarino further claim that AI regulators are at risk of turning to large tech corporations for inspiration when it comes to state-of-the art fairness practices (2023). However, this means to turn to the same companies that are shown to prefer capital gain in 3.2.1; the idea of external regulation is founded on preventing these private institutions from influencing fairness decisions to begin with. I conclude by claiming that future members of the AI Act Committee should tread carefully when using Silicon Valley corporations as inspirations, as that risks returning to square one. I close this point by saying it is important that the EU institution has systems in place to prevent overly strong Silicon Valley corporate influence.

Lastly, it is important to ensure that the regulatory institute itself does not constitute a hegemon - otherwise, the measures it imposes cannot achieve fairness representing all minority groups. Based on the concept of situated knowledge from section 3.2.2, I argue that regulations which respect the needs of as many underrepresented groups as possible can only be achieved if the regulatory institution itself is diverse. Such a group includes more perspectives on discriminatory practices, which leads to a more encompassing collective understanding of the social issues and a understanding that knowledge is situated. In addition, the institute regulating ML should interact with multiple stakeholders. Through stakeholders I on one hand mean the people directly impacted by the ML systems, so as to ensure that as many experiences as possible are accounted for. On the other hand, the limitations that the ML experts face when implementing fairness should also be investigated in order to ensure actually implementable regulations. This is especially important considering that developers have indicated difficulty implementing fairness measures, citing a disconnect between academia and real-life practices as the reason (Holstein et al., 2018).

I close these three pointers by illustrating an apparent clash between the first two critiques and the third one. The former points ask for a need to be wary of unjust corporate influence, whereas the latter highlights that interacting with ML experts from corporate contexts is important. This emphasises how difficult it is to actually get regulations such as the AI Act right when the desired outcome is social justice that accounts for everybody. One way of interacting with ML experts might be through research similar to (Holstein et al., 2018), where the interviews and companies were anonymous. By ensuring confidentiality, surveys and interviews might lead to opinions and suggestions that are not as filtered by the risks that the interviewed experts might fear facing internally.

Finally, these pointers illustrate a need for a staff that is not only diverse, but also of varying technical and legal expertise. Indeed, the need is for people with interdisciplinary backgrounds and a genuine interest in ethical implementation of technology. In the next subsection, I argue that such people are formed by our educational institutions and that this illustrates a need to rethink the way we teach ethics.

6.2 More focus should be put ethics in ML development even from the lecture halls

Following the argument from Section 3.2.2, both the AI Act and other interdisciplinary solutions require people who are diverse. However, people also need to be of varying expertise and with an interest in ethical implementation of technology. Arguing that such people are formed by our educational institutions, I showcase the need for more interdisciplinary ethics.

In their critique of technical abstraction and solutionism, Weinberg points to Malazita and Resetar's (2019) critique of computer science epistemological practices (2022). Through this, Weinberg argues that the abstraction of Computer Science leads to educating technically capable students who unfortunately lack the necessary ethical foundations to critically evaluate for whom and for what they are developing technologies. Considering this, I argue that it is very important to get ethics education right. For the scope of this paper, I analyse the ethics of TU Delft Bachelor's and Master's Computer Science and AI courses.

On one hand, the Computer Science bachelors has one mandatory ML course, teaching fundamentals of machine learning. Out of 8 weeks of content, the course staff dedicates one to Responsible Machine learning (Krijthe & Migut, 2023). In particular, the lecture raises concerns about perpetuating data biases and predictive policing, while additionally shining light on how race, for instance, is a social construct and on the complications this can bring. In particular, this course provides Johnson (2020) and Barocas et al. (2023) as materials. The former raises the concern that eliminating bias cannot be done through a purely technical solution, asking for technosolutionism. The latter takes an interdisciplinary approach by placing the technical aspects of discrimination in a societal context in chapter 8. However, this chapter is not part of the suggested reading. On the other hand, TU Delft will have a Master's course in Data Science and Artificial Intelligence starting academic edition 24/25. Out of 5 mandatory courses, one is fully dedicated to "responsible practices" in Data Engineering and Artificial Intelligence. This course will allegedly take an interdisciplinary approach to explaining unfairness in AI systems by considering both "systems and human-centric perspectives" (2024). Although not much can be concluded yet, this shows interest for interdisciplinary thinking. This is a good step when discussing ML ethics.

Next to pointing out that the problem space is not simply technical, lecturers should also dive into some concrete points of critique and solutions coming from disciplines that are not from Computer Science, but rather Social Sciences. Through this, students would gain more insight into the role that ML tools have in perpetuating or combating already existing structures of discrimination. Lastly, students would be able to question if ML solutions are indeed ethical for their use case to begin with. This idea is not important just for TU Delft, but all universities teaching ML techniques. Interacting with non-hegemonic sources of knowledge is fundamental to educating future researchers who can meaningfully listen to the needs of the most marginalised.

7 Conclusions

In this paper, I critically evaluate the current state of fairness research in the field of ML. By highlighting the significant amount of existing fairness metrics, I argue that there is too much focus on the technical details and too little on the social context. This idea has been brought forward by other authors, too, whose critiques can be summed up in three points: (1) there are too many fairness metrics, (2) the sociotechnical problem space is often reduced to a simply technical one and (3) the analysis of fairness can become too abstract, leading to incomplete analysis.

Considering this lack of societal considerations, I start contextualising ML research and algorithms from a feminist perspective. In the first argument, I use black feminist theory to illustrate that large companies are more likely to prioritise capital gain over ethical concerns. This allows me to argue that we should choose the fairness metrics which favour marginalised communities to the detriment of large-scale corporations. In the second argument, I use feminist epistemology to highlight that fairness knowledge is inherently biased, as the hegemony is predominantly composed of cisheterosexual white men. I argue that hiring more people of minority backgrounds is crucial, but that this will only bring value to fairness discussions if accompanied by a restructuring of the current corporate environments. In the last argument, I argue that critically looking at the social context and purpose of ML systems helps one conclude that some algorithms are irresponsible and unethical by definition. By using these three arguments, I create a framework which critically asks for whom, by whom and for what purpose an algorithm has been developed. I apply this framework to the COMPAS debate, showing how it permits us to see Equivant's response as a company's attempt of using technical solutionism as a tool of continuing its discriminatory practices as usual.

Finally, I use these feminist ideas to propose future solutions. Although the EU AI act is a good step towards heavier regulation of ML systems, it has already been affected by company lobbying. This problem cannot be eradicated without major political reform, but I seek hope in the idea that a more diverse regulatory body can lead to more resistance against lobbying practices, as people closer to discrimination may have a more interest in resisting the hegemony. This interest can not only be cultivated through lived experience, but through more interdisciplinary ethics education, too. Based on this, I argue that ethics classes which interact with fields such as Social Sciences can help us move towards production of knowledge with less hegemonic bias, allowing for a stronger focus on marginalised groups.

To close, I identify future research avenues. Firstly, there are people willing to advance towards more equitable ML, even within companies. Thus, more anonymous surveys that explore the needs of developers are crucial towards ensuring implementable fairness guidelines and solutions. Secondly, the EU AI Act has been shown as influenced by corporate interests, making it imperative that means of preventing unethical lobying are explored. Lastly, ethics education is fundamental to future researchers with an interest in sociotechnical fairness solutions. To ensure it is done properly, surveys evaluating the current quality of ML ethics education should be conducted.

A Identified fairness metrics

Metric	Sources
Demographic Parity	Mehrabi et al., 2019, Ferrara, 2023, Kheya et al., 2024, Ruf and Detyniecki, 2021, Barocas et al., 2023
Equal opportunity	Mehrabi et al., 2019, Kheya et al., 2024, Ferrara, 2023, Ruf and Detyniecki, 2021?, Barocas et al., 2023
Counterfactual fairness	Mehrabi et al., 2019, Ferrara, 2023, Kheya et al., 2024
Causal fairness	Ferrara, 2023, Verma and Rubin, 2017
Equalised Odds	Mehrabi et al., 2019, Verma and Rubin, 2017, Kheya et al., 2024, Ruf and Detyniecki, 2021, Barocas et al., 2023
Fairness Through (un)awareness	Mehrabi et al., 2019, Verma and Rubin, 2017, Kheya et al., 2024 Barocas et al., 2023
Treatment Equality	Mehrabi et al., 2019, Verma and Rubin, 2017, Kheya et al., 2024
Test Fairness	Mehrabi et al., 2019, Verma and Rubin, 2017
Conditional Statistical Parity	Mehrabi et al., 2019, Verma and Rubin, 2017, Kheya et al., 2024, Ruf and Detyniecki, 2021, Barocas et al., 2023
Predictive Parity	Verma and Rubin, 2017, Kheya et al., 2024, Ruf and Detyniecki, 2021, Barocas et al., 2023
Conditional Use Accuracy Equality	Verma and Rubin, 2017, Kheya et al., 2024, Ruf and Detyniecki, 2021, Barocas et al., 2023
Balance for Positive/Negative class	Verma and Rubin, 2017, Kheya et al., 2024, Ruf and Detyniecki, 2021, Barocas et al., 2023
Fairness through (un)awareness	Verma and Rubin, 2017, Kheya et al., 2024
Unresolved discrimination	Verma and Rubin, 2017, Kheya et al., 2024
Predictive Equality	Kheya et al., 2024, Ruf and Detyniecki, 2021, Barocas et al., 2023
Calibration (within groups)	Verma and Rubin, 2017, Barocas et al., 2023, Kheya et al., 2024, Ruf and Detyniecki, 2021
Group Fairness	Barocas et al., 2023, Verma and Rubin, 2017

Table 1: Fairness metrics

References

(2024). https://www.tudelft.nl/onderwijs/opleidingen/masters/dsait/msc-data-science-and-artificial-intelligence-technology/programme

Acker, J. (2006). Inequality regimes: Gender, class, and race in organizations. *Gender &amp; Society*, *20*(4), 441–464. https://doi.org/10.1177/0891243206289499

Allen, A. (2022). Feminist perspectives on power (E. N. Zalta & U. Nodelman, Eds.) [Backup Publisher: Stanford Encyclopedia of Philosophy Edition: Fall 2022]. https://plato.stanford.edu/entries/feminist-power/#PoweDomi

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May). Machine bias. Retrieved May 1, 2024, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press.

Bates, T. R. (1975). Gramsci and the theory of hegemony. *J. Hist. Ideas*, *36*(2), 351.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bernhagen, P., & Bräuninger, T. (2005). Structural power and public policy: A signaling model of business lobbying in democratic capitalism. *Political Studies*, *53*(1), 43–64. https://doi.org/10.1111/j.1467-9248.2005.00516.x

Bland, M. (2020, July). Algorithms can predict domestic abuse, but should we let them? https://doi.org/10.1007/978-3-030-50613-1_6

Caglayan, A., Horsanali, M. O., Kocadurdu, K., Ismailoglu, E., & Guneyli, S. (2022). Deep learning model-assisted detection of kidney stones on computed tomography. *International braz j urol*, *48*(5), 830–839. https://doi.org/10.1590/s1677-5538.ibju.2022.0132

Carlton, B. (2016). Penal reform, anti-carceral feminist campaigns and the politics of change in women's prisons, victoria, australia. *Punishment & Society*, *20*(3), 283–307. https://doi.org/10.1177/1462474516680205

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv:1808.00023 [cs]*. https://arxiv.org/abs/1808.00023

Dakalbab, F., Abu Talib, M., Abu Waraga, O., Bou Nassif, A., Abbas, S., & Nasir, Q. (2022). Artificial intelligence &amp; crime prediction: A systematic literature review. *Social Sciences &amp; Humanities Open*, *6*(1), 100342. https://doi.org/10.1016/j.ssaho.2022.100342

Dastin, J. (2018, October). Insight - amazon scraps secret AI recruiting tool that showed bias against women (W. Jonathan & M. Dickerson, Eds.) [Backup Publisher: Reuters]. Retrieved May 1, 2024, from https://www.reuters.com/article/idUSKCN1MK0AG/

Davis, A. (2000). *Indigenous Law Bulletin*, *4*(27), 4–7. https://search.informit.org/doi/10.3316/ielapa.200009064

Demers, J. (2013). Control, resistance, and racism in the contemporary prison. *Can. Rev. Am. Stud.*, *43*(3), 502–511.

Dieterich, W., Mendonza, C., & Brennan, T. (2016). Compas risk scales : Demonstrating accuracy equity and predictive parity performance of the compas risk scales in broward county. https://api.semanticscholar.org/CorpusID:51920414

Draude, C., Klumbyte, G., Lücking, P., & Treusch, P. (2019). Situated algorithms: A sociotechnical systemic approach to bias. *Online Inf. Rev.*, *44*(2), 325–342.

Equivant. (2020, January). https://www.equivant.com/

Ferrara, E. (2023). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *arXiv:2304.07683 [cs]*. https://arxiv.org/abs/2304.07683

Fisher, B., Tronto, J., Abel, E. K., & Nelson, M. (1990). Toward a feminist theory of caring. *Circles of care*, 29–42.

Fraser, H., & Bello y Villarino, J.-M. (2023). Acceptable risks in europe's proposed ai act: Reasonableness and other principles for deciding how much risk management is enough. *European Journal of Risk Regulation*, 1–16. https://doi.org/10.1017/err.2023.57

Gebru, T. (2020a, July). Race and gender (M. D. Dubber, F. Pasquale, & S. Das, Eds.). https://doi.org/10.1093/oxfordhb/9780190067397.013.16

Gebru, T. (2020b, December). https://x.com/timnitGebru/status/1334352694664957952

George, A., A.S.Hovan George, & A.S.Gabrio Martin. (2023). The environmental impact of ai: A case study of water consumption by chat gpt. https://doi.org/10.5281/ZENODO.7855594

Gonzalez, M. F., Liu, W., Shirase, L., Tomczak, D. L., Lobbe, C. E., Justenhoven, R., & Martin, N. R. (2022). Allying with ai? reactions toward human-based, ai/ml-based, and augmented hiring processes. *Computers in Human Behavior*, *130*, 107179. https://doi.org/10.1016/j.chb.2022.107179

Grasswick, H. (2018). Feminist Social Epistemology. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2018). Metaphysics Research Lab, Stanford University.

Green, B., & Hu, L. (2018). *The myth in the methodology: Towards a recontextualization of fairness in machine learning*. https://api.semanticscholar.org/CorpusID:49544563

Hampton, L. M. (2021). Black feminist musings on algorithmic oppression. *CoRR*, *abs/2101.09869*. https://arxiv.org/abs/2101.09869

Holstein, K., Vaughan, J. W., Daumé, H., Dudík, M., & Wallach, H. (2018). Improving fairness in machine learning systems: What do industry practitioners need? https://doi.org/10.48550/ARXIV.1812.05239

Jassy, A. (2024). Annual report 2023: Letter to shareholders [[Accessed 23-06-2024]]. https://ir.aboutamazon.com/annual-reports-proxies-and-shareholder-letters/default.aspx

John-Mathews, J.-M., Cardon, D., & Balagué, C. (2022). From reality to world. a critical perspective on ai fairness. *Journal of Business Ethics*, *178*(4), 945–959. https://doi.org/10.1007/s10551-022-05055-8

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

Kheya, T. A., Bouadjenek, M. R., & Aryal, S. (2024). The pursuit of fairness in artificial intelligence models: A survey [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2403.17333

Krijthe, J. H., & Migut, M. A. (2023). https://brightspace.tudelft.nl/d2l/le/content/595332/Home

Lauderdale, P., & Amster, R. (2008). Power and deviance. In *Encyclopedia of violence, peace, &amp; conflict* (pp. 1696–1703). Elsevier. https://doi.org/10.1016/b978-012373985-8.00143-4

Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376445

Mahler, T. (2022). Between risk management and proportionality: The risk-based approach in the eu's artificial intelligence act proposal. *The Swedish Law and Informatics Research Institute*, 247–270. https://doi.org/10.53292/208f5901.38a67238

Majumder, S., Chakraborty, J., Bai, G. R., Stolee, K. T., & Menzies, T. (2023). Fair enough: Searching for sufficient measures of fairness [Publisher: Association for Computing Machinery]. *ACM Transactions on Software Engineering and Methodology*. https://doi.org/10.1145/3585006

Malazita, J. W., & Resetar, K. (2019). Infrastructures of abstraction: How computer science education produces anti-political subjects. *Digit. Creat.*, *30*(4), 300–312.

Mammonas, D. (2024, May). Artificial intelligence (ai) act: Council gives final green light to the first worldwide rules on ai. https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/

McInerney, K. (2023, October). Coding 'carnal knowledge' into carceral systems: A feminist abolitionist approach to predictive policing. In *Feminist ai* (pp. 101–118). Oxford University PressOxford. https://doi.org/10.1093/oso/9780192889898.003.0007

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv:1908.09635 [cs]*. https://arxiv.org/abs/1908.09635

Nadella, S. (2023). Annual report 2023: Shareholder letter [[Accessed 23-06-2024]]. https://www.microsoft.com/investor/reports/ar23/

Noble, S. U. (2020, May). *Algorithms of oppression: How search engines reinforce racism*. New York University Press. https://doi.org/10.18574/nyu/9781479833641.001.0001

Phan, T., Goldenfein, J., Mann, M., & Kuch, D. (2021). Economies of virtue: The circulation of 'ethics' in big tech. *Science as Culture*, *31*(1), 121–135. https://doi.org/10.1080/09505431.2021.1990875

ProPublica. (2017, February). https://www.propublica.org/about/

Richardson, B., & Gilbert, J. E. (2021). A framework for fairness: A systematic review of existing fair AI solutions [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2112.05700

Roberts, A. (2014). The political economy of "transnational business feminism": Problematizing the corporate-led gender equality agenda. *International Feminist Journal of Politics*, *17*(2), 209–231. https://doi.org/10.1080/14616742.2013.849968

Rozado, D. (2023). The political biases of chatgpt. *Social Sciences*, *12*(3). https://doi.org/10.3390/socsci12030148

Rudin, C., Wang, C., & Coker, B. (2020). The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, *2*(1). https://doi.org/10.1162/99608f92.6ed64b30

Ruf, B., & Detyniecki, M. (2021). Towards the right kind of fairness in ai.

Salman, C. (2024). Cool it! the objective racism of carceral technofixes. *Critical Studies in Media Communication*, *41*(1), 21–35. https://doi.org/10.1080/15295036.2024.2314667

Saltelli, A., Dankel, D. J., Di Fiore, M., Holland, N., & Pigeon, M. (2022). Science, the endless frontier of regulatory capture. *Futures*, *135*(102860), 102860.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency.*

Singer, N. (2019). *Amazon is pushing facial technology that a study says could be biased.* https://www.nytimes.com/2019/01/24/technology/amazon-facial-technology-study.html

Sundar, P. (2024). Form 10-k for the fiscal year ended december 31, 2023 [[Accessed 23-06-2024]]. https://abc.xyz/investor/

Tronto, J. C. (1998). An ethic of care. *Generations: Journal of the American Society on Aging*, *22*(3), 15–20. Retrieved June 23, 2024, from http://www.jstor.org/stable/44875693

Tuck, E., & Yang, K. W. (2014). Unbecoming claims: Pedagogies of refusal in qualitative research. *Qualitative Inquiry*, *20*(6), 811–818. https://doi.org/10.1177/1077800414530265

Veale, M., & Borgesius, F. Z. (2021). *Computer Law Review International*, *22*(4), 97–112. https://doi.org/doi:10.9785/cri-2021-220402

Verma, S., & Rubin, J. (2017). Fairness definitions explained. https://doi.org/10.1145/3194770.3194776 von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust ai. *Philosophy &amp; Technology*, *34*(4), 1607–1622. https://doi.org/10.1007/s13347-021-00477-0

Vyas, D. A., Eisenstein, L. G., & Jones, D. S. (2020). Hidden in plain sight — reconsidering the use of race correction in clinical algorithms (D. Malina, Ed.). *New England Journal of Medicine*, *383*. https://doi.org/10.1056/nejmms2004740

Waldron, J. (1993, March). *Cambridge studies in philosophy and public policy: Liberal rights: Collected papers 1981-1991*. Cambridge University Press.

Weinberg, L. (2022). Rethinking fairness: An interdisciplinary survey of critiques of hegemonic ML fairness approaches. *Journal of Artificial Intelligence Research*, *74*, 75–109. https://doi.org/10.1613/jair.1.13196

West, M., Sarah, Whittaker, M., & Crawford, K. (2019). *Discriminating systems: Gender, race and power in ai* (tech. rep.). AI Now Institute. https://ainowinstitute.org/discriminatingsystems.html

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019). Explainable ai: A brief survey on history, research areas, approaches and challenges. In *Lecture notes in computer science* (pp. 563–574). Springer International Publishing. https://doi.org/10.1007/978-3-030-32236-6_51

Yakinova, Y., & Ojamo, J. (2024). Artificial intelligence act: Meps adopt landmark law [[Accessed 23-06-2024]].

Zou, L., & Khern-am-nuai, W. (2022). Ai and housing discrimination: The case of mortgage applications. *AI and Ethics*, *3*(4), 1271–1281. https://doi.org/10.1007/s43681-022-00234-9