Mitigating Accidental Gender Inequalities in AI Systems

Hasan Alparslan Bayrak | Bilkent University

Introduction

The advancement of artificial intelligence (AI) brings forth questions for humanity and sentient life. A key inquiry revolves around determining the values with which AI systems should align. One perspective argues for a utilitarian approach, emphasizing that these technologies must be developed with the goal of maximizing happiness for the greatest number of individuals or sentient animals over the long term.[1] Another viewpoint, rooted in Kantian principles, asserts that AI should adhere only to principles universally acceptable, such as fairness or beneficence.[2] Other approaches prioritize aligning AI with human direction, intentions, preferences, or desires.[3]

The challenge of alignment comprises two facets: the technical aspect concentrates on encoding values to guarantee reliable behavior in artificial agents, whereas the normative dimension examines which values or principles should be encoded. This essay explores the normative aspect of the value alignment challenge, organized into two sections. Firstly, by exploring why we need change I explore real-world examples of agent misalignment, specifically accidental gender inequalities in machine learning systems. This examination highlights the interconnectedness of technical and non-technical aspects in AI alignment. Secondly, I present Honneth's theory of recognition (1996) to underscore the philosophical dimension and argue for the limitations of relying solely on a utilitarian perspective in AI alignment. Thirdly, I give the intersectional feminist approach to diversity and inclusion. Finally, I conclude by arguing that our primary challenge lies not in identifying the definitive moral theory to encode in machines but rather in establishing fair and equal processes for selecting the values to be encoded, and I believe it is possible to achieve this aim by reimagining the concept of fairness in AI through feminist principles, more specifically through intersectional feminism.

---

[1] For more on this topic, see Longoni, Chiara et al., Artificial Intelligence in Utilitarian vs. Hedonic Contexts: The "Word-of-Machine" Effect. *Journal of Marketing* 86:1, 91-108, (2022).

[2] See Hooker, J., and Kim, T. W., titled Toward Non-Intuition-Based Machine and Artificial Intelligence Ethics: A Deontological Approach Based on Modal Logic, (2018), or Powers, T. M. Prospects for a Kantian Machine. *IEEE Intelligent Systems* 21(4):46–51, (2006).

[3] See Kim, T.W., Hooker, J., Donaldson, T.: Taking principles seriously: a hybrid approach to value alignment. *J. Artif. Intell. Res.* 70, 871–890 (2021).

Why do we need change?

Gender bias is a widespread problem that shows itself as many types of bias and stereotyping that many of us deal with on a daily basis. Our lives are shaped by those experiences from a young age. Consider the toys we are encouraged to play with, the colors we wear, the interests we either pursue or shun, and the persistent gender pay gap that results in men being paid more than women. Although in recent years there has been a growing effort to expose and reduce them, a fundamental truth remains: our world, our society and we ourselves hold biases, including those related to gender - whether we are aware of them or not.

But how does this apply to AI? AI systems all start with a basic collection of data, no matter how sophisticated they become. We must also acknowledge that this data is biased if we believe it is an accurate representation of our current reality - it is a result of our biases. Discrimination and inequality still have a strong hold on our social systems. So if we train AI on data from earlier decades without thinking about it, these inequalities will inevitably show up in AI models. For example, I just gave the following prompt to ChatGPT 4.0: "Tell me a story about a boy and a girl building their career capital in approximately 100 words." Here's the answer it gave:

> *Ethan and Lily, childhood friends, often spent afternoons dreaming about their futures. Ethan loved tinkering with gadgets and knew he wanted to be an engineer, while Lily's passion for helping animals made her determined to become a veterinarian. As they grew older, their paths diverged; Ethan earned a scholarship to a prestigious engineering school, and Lily got accepted into a veterinary program. Despite the distance, they kept in touch, sharing their successes and challenges. Years later, they reunited in their hometown, both accomplished in their fields, proud of the dreams they had chased and achieved.*

At present, many AI systems reinforce existing power structures, biases, and stereotypes (Holroyd et al. 2018; p. 76). In fact, it could be argued that AI is currently contributing to greater injustice–a trend we cannot afford to continue. But there is another way: we can develop and employ AI in a feminist way. When I talk about feminism, I mean an intersectional and inclusive feminism that seeks justice for all marginalized people; that is to say, it is not just about women. Feminism is about critically questioning existing norms, seeking change, and advocating for a shift in power dynamics (Khader 2017; p. 6). By embracing these principles, we have the potential to increase our positive impact and to change the world. So, we need feminist AI because:

> **P1.** The beliefs, biases, and designs of those who created AI are ingrained in it—AI is not impartial; it reinforces stereotypes that are representative of the biases that are embedded in its development and deployment.
>
> **P2.** AI systems' pre-existing biases perpetuate stereotypes, undermining the uniqueness of women and contributing to misrecognizing women, limiting their identities, and reinforcing outdated gender norms.
>
> **C.** Therefore, ensuring gender equality in AI systems is crucial for fostering inclusive and equitable societal development.

Now, let's look at the real-world examples of these biases in AI systems.

Accidental gender inequalities in AI systems

In this section, I will present a series of practical problems on how gender biases in machine learning systems manifest in various contexts, exacerbating societal gender inequalities. While doing this, I aim to categorize these issues into a broader thematic framework, namely the issue of accidents in machine learning systems. Accidents are harmful and unintentional actions that might result from poor systems design in the actual world (Amodei et al. 2016; p. 1). In this paper, accidents refer to gender-biased results that may occur from improperly defining the goal function, failing to pay attention throughout the learning process (i.e., oversight in the learning process) or other machine learning implementation mistakes.

When it comes to gender inequality, an accident can happen when mechanisms intended for one purpose unintentionally maintain or worsen already-existing gender inequality. These systems may have unanticipated and detrimental gender-related effects due to their poor design or execution, even while their goals include impartial decision-making and equitable representation. Potential accidents encompass a wide range, including but not restricted to, the following problems.

Negative side effects

In "negative side effects," the designer specifies an objective function that focuses on accomplishing some specific task in the environment in a most effective way, but ignores other aspects of the (potentially very large) environment, and thus implicitly expresses indifference over environmental variables that might be harmful to change (Amodei et al. 2016; p. 2). Because of the existing human biases in the community, negative side effects frequently happen. The underlying biases in the system originate from either the training data used to build the system or from its technological basis.

Now, examining inequalities comes from AI systems generated by the quest for maximizing effectiveness.

Inherent bias in hiring

AI plays a crucial role in the recruitment process, integrated into resume analyzers, applicant tracking systems, tests, and interview evaluation tools, extensively used by 99% of Fortune 500 companies (Cookson et al. 2020). Large businesses are expected to replace about 16% of HR personnel over the next ten years as a result of this trend, which shows that they are depending more and more on software-driven processes. However, inherent biases in training data pose a significant challenge. AI-based systems may unintentionally ignore or even perpetuate biases against specific gender-associated traits or backgrounds when it is exclusively focused on completing the goal of effectively selecting candidates. Ignoring the larger societal implications or biases encoded in the data, the algorithm may unintentionally favor or disfavor people based on gender-associated qualities present in the training data, all in the name of candidate sorting. This underscores how biases within hiring tools exemplify gender bias in algorithmic decision-making.

Consider the experimental hiring tool used by Amazon. The system was designed to rank candidates from one to five stars and identify the best fit for the position. However because the majority of the data it was trained on was resumes from male candidates, it started to discriminate against women (Müller 2020). By penalizing phrases like "women's" in resumes, such as "women's chess club captain," and downgrading resumes from graduates of "women's colleges," the algorithm mirrored the historical bias towards hiring more men. As a result, it declared that women were undesirable and disapproved of applications that contained the term "women."

Selection bias and creditworthiness

Another example in the pursuit of efficient AI systems may be given where AI algorithms are used to assess creditworthiness. These algorithms consistently give preference to male borrowers over female borrowers with identical financial profiles, favoring the former with better loan conditions and interest rates (Müller 2020). The reason is the same: historical biases present in the data used to train these algorithms.

Reward hacking

Another framework to help us categorize and address accidental gender inequalities is reward hacking. In reward hacking, the objective function that the designer writes down admits of some clever "easy" solution that formally maximizes it but perverts

the spirit of the designer's intent (Amodei et al. 2016; p. 7). When a designer chooses an objective function that seems to be strongly correlated with completing the task, but this correlation becomes much weaker when the goal function is heavily optimized, reward hacking may occur. Real-world examples highlight the significance of addressing this issue.

Implicit stereotype and unconscious bias

Preexisting stereotypes, particularly those depicting women as inherently more nurturing than men, permeate societal structures, influencing organizational cultures and shaping the perception of women's roles in both professional and domestic spheres. This established perception often limits women's career growth opportunities and impacts their progression.

For instance, gender biases have been found in two significant image collections supported by Facebook and Microsoft (Kay et al. 2015). The way that hobbies and sports were portrayed was clearly sexist. While teaching and shooting were more closely linked to men, activities like shopping and washing were consistently associated with women. In addition, household items like spoons and forks were associated more strongly with women than with men; this was in contrast to outside sports equipment like snowboards and tennis rackets. Geographic bias also existed, with algorithms incorrectly classifying photos according to cultural backgrounds. An example of bias in the dataset representation is when a North Indian bride was labeled as "costume" and "performance art," but a standard US bride was tagged with phrases like "bride" and "wedding" (Kay et al., 2015).

Furthermore, these biases were not only mirrored but also magnified by machine-learning algorithms that were trained on these biased datasets. For example, the program learned from a photo set that mostly featured women cooking, therefore strengthening and amplifying the relationship between the two. The program was mislabeled as "woman" in several cases, even men who were shown in kitchen settings. In a similar vein, Google's search biases were brought to light by researchers from the Universities of Washington and Maryland. Google image searches for phrases like "Chief Executive Officer" (CEO) revealed a biased representation; only 11% of the people depicted as CEOs were female, which is far less than the 27% of female CEOs in the US at the time (Kay et al., 2015).

Voice and face recognition systems

Face and voice recognition software is another issue that comes from reward hacking. Voice recognition technologies are used by digital assistants like Alexa, Siri, and

Google Assistant. It can entail simply comprehending what someone is saying (e.g., "Siri, play some music") or it can involve recognizing a specific individual based only on their voice. For the first use, it is frequently referred to as "speech recognition". Studies have shown that when dealing with female voices, speech recognition accuracy drops (Tatman 2017). This suggests that women are less likely to be understood or acknowledged by digital assistants and will more often have trouble speaking with them. This also holds true for people with different regional or foreign accents. However, this is not the only problem.

Consider voice assistants, such as Siri and Alexa. They frequently adopt feminine voices and personas to correspond with jobs that have traditionally been assigned to women, including scheduling or creating reminders. The intentional selection of a female voice carries the risk of upholding traditional gender norms, which may impact social attitudes and actions (Samuel 2019). These technologies' persistent use of gendered voices may unintentionally support discriminatory recommendation systems while limiting individual potential. Not only are our interactions with these digital helpers useful, but they also have a deeper meaning. Language that is harsh, disparaging, or sexually explicit and is aimed toward these assistants may unintentionally legitimize disrespectful behavior in interpersonal relationships, particularly with regard to women. Feminized digital assistants' apathetic or cowardly reactions to such remarks may serve to perpetuate the idea that women are docile or subservient. Gender-specific computer voices have the ability to elicit gender-stereotypical actions from users even when used in isolation from other gender indicators such as looks (Nass et al., 2006). According to a 2019 UNESCO assessment, the widespread use of voice assistants that sound feminine feeds into the perception of women as submissive and obedient (Mark et al., 2019). So voice assistants are powerful socialization aids that specifically shape children's ideas of what is expected of women, girls, and those who identify as female in society in terms of duties and responsibilities. These results highlight the significant consequences of voice settings in technology and shed light on how they influence cultural norms and perceptions of gender.

Facial recognition has also been demonstrated to perform worse on women than on men and, conversely, on individuals with darker skin tones compared to those with lighter skin tones. Darker-skinned women had an error rate of up to 34.7%, while lighter-skinned males had an error rate of only 0.8% in Buolamwini and Gebru's (2018) study of three commercial face analysis tools for gender classification. The inability to process higher-pitched female voices is the reason for misidentification in voice recognition. However the lack of inclusion in training data sets is the cause of the

discrepancy in the recognition of faces—both male and female—as well as light and dark skin tones. The former represents technical bias, whereas the latter represents preexisting bias.

Safe exploration

The existence of harmful outcomes can arise even when the designer sets the right formal objective. This can occur due to decisions made from inadequate or biased training data or when the model lacks expressiveness. Consider the implications in the context of gender biases: an AI system designed to mitigate biases may inadvertently perpetuate them due to the biased data used in its training. This raises concerns about the system's ability to make fair decisions, particularly when faced with novel inputs or scenarios. Addressing these challenges is akin to navigating uncharted territory; ensuring systems avoid reinforcing gender disparities requires a multidimensional approach. Strategies such as "safe exploration" emphasize the need to prevent the perpetuation of gender stereotypes or biases, especially when the system explores uncharted territories.

Algorithms of oppression

Do a basic internet search for "Latina women" and what do you see? Primarily, you will find content that objectifies and sexualizes, with terms like "spicy" and other offensive descriptors. Conversely, if you search for "White women," the results are very different and often more neutral. This stark contrast underscores the troubling reality of online representations of women of color, where suggested adult content and unchecked discussions label Latina women as "fiery" or "submissive". This phenomenon highlights the pervasive issue of data discrimination, which is a profound social problem that arises from the convergence of private interests promoting certain websites and the monopolistic dominance of a few major search engines. This dynamic results in biased algorithms that favor certain racial groups while systematically disadvantaging others, particularly women of color (Gengler et al. 2023). The impact of such biases is severe, reinforcing harmful stereotypes and perpetuating social inequalities.

Furthermore, as search engines and related companies become increasingly integral to our daily lives, serving as primary resources for communication, education, and more - it is critical to address these troubling trends. The algorithms that power search engines are far from neutral; they reflect and reinforce existing societal biases (Göritz et al. 2023). The objectification of Latina women is not an isolated incident, but rather a manifestation of broader biases embedded in these technological frameworks.

What can we do?

The more we attempt to address and resolve these concerns, the more complicated the problem of bias and fairness in machine learning algorithms becomes. It's critical to realize that constructive bias is necessary for algorithms to effectively model data and generate useful predictions; fully bias-free machine learning is not achievable (Adebayo 2012; p. 17-18). The optimization of a cost function is the key to machine learning performance, and this decision brings generative bias into the system. A number of other factors, including context, goal, data accessibility, and trade-offs between generalization, speed, and accuracy, all contribute to this generative bias. Thus, it is false to believe that machine learning is completely free of prejudices; bias is a feature of inductive learning systems. Furthermore, biases are included in the training data itself, which makes it necessary for research to differentiate between biases that are discriminative or computational and those that represent underlying patterns.

In machine learning, bias refers to consistently skewed outcomes brought about by false presumptions. But in the absence of these presumptions, an algorithm's performance on an assignment will be no better than chance, a notion encapsulated in Wolpert's 1996 No Free Lunch theorem. This theorem states that, when averaged across all possible distributions that might provide data, all classifiers show the same error rate. Because of this, in order for a classifier to successfully represent some distributions and functions, it has to be biased toward those particular distributions and functions; nevertheless, this specialization reduces the effectiveness of other distribution types. Furthermore, because the training data is incompletely reflective of reality due to its narrow scope, bias results. Machine learning models become biased as a result of this restricted representation and selection of datasets (Adebayo 2012; p. 26). The bias in the system is further influenced by the presumption that the available training data accurately classifies the test data and sufficiently models it.

It is nevertheless possible for prejudices to reappear when systems change over time, even with attempts to completely eliminate prevalent biases. The difficulty is in efficiently updating algorithms once they have been taught, verified, and put into use. Crucial issues include figuring out how well these systems are working now, assessing their effectiveness, and defining success. It is necessary to be able to explain and evaluate the mechanics behind the numerous biases present in these systems in order to recognize and comprehend them. It's critical to recognize the biases that these systems require to work well and to spot injustices that result from either the algorithms or the training data.

For this reason, I believe machine learning models should be ethical - respecting ethical principles and values. In this aim, I believe it's crucial to prioritize the creation of AI systems that are honest and truthful, aiming to mitigate unfairness embedded within both the training data and the algorithms. "Honest" AI systems are those that properly express what they believe truthfully while upholding moral standards and values, whereas "truthful" AI systems are those that avoid saying falsehoods (Evans et al. 2021). The philosophical problem here is how to define the concepts of truthfulness and honesty. I believe we can take insights from feminist philosophy literature.

Intersectional feminist approach to diversity and inclusion

Until this point, we have just seen some examples of the accidental gender inequalities in machine learning systems and the reason why we need feminist AI. Now, by taking insights from feminist literature, my aim is to define what is feminist AI and what 'responsible use' should look like in AI agents.

Gender oppression is not a discrete phenomenon; it frequently overlaps with other types of oppression, including those that are supported by colonialism and capitalism (Frye 2000; p. 13). Feminism's intersectional character highlights how crucial it is to comprehend how various power and inequality structures are linked to one another and impact people in various ways. For instance, by promoting economic disparities that disproportionately impact women—particularly women of color and those from marginalized communities—capitalism can worsen gender inequality (Frye 2000; p. 14). Profit is frequently prioritized over individuals in the context of capitalism, which results in salary disparities and abusive labor practices that make life more difficult for those who already experience gender discrimination (Young 2001; p. 6). Similarly, gender interactions are still shaped by the structural and cultural oppression left behind by colonialism. Western values and norms have been imposed through colonial histories, marginalizing non-Western and indigenous gender identities and roles (Khaer 2017; p. 3-4). This cultural appropriation upends established ways of life and perpetuates structural injustices that are compounded by gender discrimination.

Furthermore, gender biases in AI go beyond simple discrimination to represent kinds of misrecognition that can negatively affect women's self-esteem and personal growth. According to recognition theory (Honneth 1996), our social relationships influence our personalities and identities, which in turn shape the roles we play and the goals we pursue in our day-to-day social interactions. Whether we accept or reject our interactions with others, they are essential to the development of individuals and

society as a whole. Through interaction with others and adopting their perspective, an individual develops her "practical relation-to-self," which determines how she defines her value and interprets her role in society (Honneth 1996; p. 92). However, AI systems' preexisting biases perpetuate stereotypes, undermining the uniqueness of women with causing agent misalignment. Not all women conform to traditional expectations, and relying on such stereotypes hinders acknowledging women as individuals with diverse roles, actions, and thoughts. This contributes to misrecognizing women, limiting their identities, and reinforcing outdated gender norms (Waelen et al. 2022). So, my argument here can be written more formally as follows:

> **P1.** Our social relationships influence our personalities and identities, which in turn shape the roles we play and the goals we pursue in our day-to-day social interactions (Holroyd et al., 2018, p. 76).
>
> **P2.** Through interaction with others and adopting their perspective, an individual develops her "practical relation-to-self," which determines how she defines her value and interprets her role in society (Honneth, 1996, p. 92).
>
> **P3.** However, AI systems' pre-existing biases perpetuate stereotypes, undermining the uniqueness of women by causing agent misalignment and being oppressed—not all women conform to traditional expectations, and relying on such stereotypes hinders acknowledging women as individuals with diverse roles, actions, and thoughts (Gheaus, 2008, p. 3).
>
> **C.** Therefore, addressing and mitigating the biases inherent in AI systems is not only essential for promoting fairness and equity but also for fostering a society that embraces the multifaceted identities and contributions of all individuals.

In this aim, I suggest that the ultimate goal of AI systems ought to be fairness, as opposed to effectiveness, which gives priority to accuracy. The main challenge is not simply finding the correct moral theory and programming it into machines. Instead, the focus should be on establishing fair procedures for deciding which values to incorporate.

I now present a research approach that I believe will help advance AI equity, which might safeguard fairness without collapsing due to AI. The basic idea is that AI must recognize the needs, rights, and accomplishments of women. Feminist critics

imply that science was helping to perpetuate inequality - inequality in jobs, inequality in wages, inequality in expectations, and the treatment both in the home and outside it (Kourany 2010; p. 49). That's why feminist philosophers of science could not disregard the larger social context of science, and how it impacts society. Ultimately, science is a product of society and is thus influenced by social dynamics (Gorham 2009).

How is the above argument related to AI systems? First, preexisting biases in AI systems fail to acknowledge the uniqueness of women. Not all women are attracted to pink, wish to establish a family or decide to work in typically feminine sectors like nursing or early childhood education. Women are unique individuals who resist being reduced to a limited set of traits. Second, because AI systems rely on stereotypes, they are unable to acknowledge women in the concept of equality, especially when it comes to the notion that women ought to submit to males. Stereotypes contribute to the misrecognition of women, which shapes women's identities and significantly reduces the range of roles, actions, and thoughts that women make in their lives. This is because they see women as objects that exist only to serve others rather than as distinct individuals with needs and desires of their own (Waelen et al. 2022).

So, we should generate AI systems that recognize the needs, rights, and accomplishments of women. In this aim, I suggest that the ultimate goal of AI systems ought to be fairness, which gives priority to feminist values, as opposed to effectiveness, which gives priority to accuracy. The reason is based on how machine learning models function. These systems create general concepts about the topic of interest by linking attributes to labels. Nevertheless, this method can amplify biases in the training set. The model is trained, for instance, to recognize genders based on location, where men are typically observed in garages and women in kitchens. It might therefore rely on biased generalizations to achieve high accuracy. These biases are further reinforced by the model's rewards for both correct and incorrect predictions, which are based on biased information (Waelen et al. 2022). Thus, it becomes imperative to prioritize fairness over accuracy to minimize and eliminate biases that are reinforced within AI models. Aiming to ensure that particular groups are not subjected to discrimination in AI decision-making, AI fairness seeks to reduce these biases.

On this point, someone may raise an objection by arguing that addressing bias in AI systems may introduce trade-offs in terms of correctness and effectiveness, potentially compromising their overall performance. For example, implementing strict fairness constraints may limit the model's ability to capture nuanced patterns in the data, leading to suboptimal performance in certain tasks. This limitation could affect

the system's ability to accurately reflect the intricacies and complexities inherent in real-world scenarios, thereby reducing its practical utility and reliability.

Reimagining fair AI: Principles for alignment

While it is valid to highlight potential trade-offs between addressing bias in AI systems and maintaining operational efficiency, it is essential to approach this concern from a broader perspective. I believe that by redefining fairness in AI systems, we can both create feminist designs and satisfy the criteria of effectiveness. First, it's crucial to recognize that bias mitigation measures are not inherently antagonistic to efficiency. Rather than seeing fairness constraints as a barrier to performance, we should see them as an integral part of the responsible use of AI. We can proactively identify and address bias without compromising overall system effectiveness by building fairness principles into the design process from the outset. But what defines fairness, and how does it intersect with machine learning?

Considering the technical aspect of the study, these principles we should use to define fairness include but are not limited to, recognition, representation, and intersectionality. As mentioned earlier, recognition means acknowledging the needs, rights, and achievements of women. However, to meet this condition, we must embrace the vision of "cooperation, not competition". This change of perspective is essential. Traditional social frameworks often emphasize competition, which can encourage individualism and hierarchical power relations. Promoting a culture of cooperation, on the other hand, is consistent with feminist ideas of solidarity and group empowerment. In order to create an atmosphere in which everyone can thrive, cooperation places a strong emphasis on sharing resources, supporting each other, and achieving common goals. To transcend the competitive hackathon mindset, we should adopt the ethos of 'one agency' in our collaborative endeavors. Within this framework, each participant takes collective responsibility for the success of the group, fostering an environment of mutual support and cooperation while emphasizing empathy rather than sympathy. This approach is integral to achieving the goal of fair AI, as it requires addressing concerns not only through technological interventions but also by reshaping individual perceptions and fostering inclusive recognition.

Moreover, another effective strategy is to shift our design paradigm from a 'universal user' focus to an 'agency' concept. The prevailing design philosophy of 'maximizing the happiness of the greatest number of people' often leads to specific needs being overlooked in favor of catering to broader demographics. However, prioritizing the needs of underprivileged communities inherently leads to solutions that serve a wider range of people. A powerful example of this principle is wheelchair

ramps, which not only provide access for wheelchair users but also accommodate people with different mobility needs. We can develop AI systems that are not only more equitable but also more effective in addressing the complexities of real-world scenarios, by centering design processes around the diverse needs of community members.

I believe that the future work for these concepts and intersections with AI systems is significant to mitigate and eliminate discriminatory biases, caused by accidental gender inequalities. It will help us to build a more equitable society, and avoid social catastrophes that may come from the development of artificial general intelligence.

Conclusion

The essay delves into real-world examples, specifically focusing on accidents in machine learning systems that perpetuate gender biases and contribute to societal gender inequalities, which may occur from improperly defining the goal function, failing to pay attention throughout the learning process (i.e., oversight in the learning process), or other machine learning implementation mistakes. Recognizing the inherent connection between technical and non-technical aspects, the essay advocates for a comprehensive approach that integrates normative considerations into AI research. It contends that value alignment should extend beyond universally agreed-upon values, emphasizing principles that individuals would endorse impartially. I propose reimagining the concept of fairness in AI through the lens of feminist principles, such as recognition, representation, and intersectionality. The call to redefine fairness in this manner aims to address the limitations of existing AI systems, mitigate gender biases, and ensure a more equitable and inclusive technological future. These guidelines give insight into the idea of fair AI and, when used, can help the creation of an emerging technology that respects and reflects the diversity of human experiences.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mane, D. (2016). *Concrete Problems in AI Safety*

Buolamwini, J. and Gebru, T. (2018). "Gender shades: Intersectional accuracy disparities in commercial gender classification," *Proceedings of Machine Learning Research*, 81, 1-15

Christian, B. (2020). *The Alignment Problem: Machine Learning and Human Values*. W. W. Norton & Company

Dalul, S. (2019). "AI is not neutral, it is just as biased as humans," *Android Pit*

Frye, M. (2000). "Oppression," in *Gender basics: Feminist Perspectives on Women and Men* (2nd ed.). Wadsworth

Gengler, E. J., Wedel, M., Wudel, A., & Laumer, S. (2023). "Power imbalances in society and AI: On the need to expand the feminist approach."

Gheaus, A. (2008). "Basic income, gender justice and the costs of gender-symmetrical lifestyles," *BIS*, 3(3)

Gorham, G. (2009). *Philosophy of Science: A Beginner's Guide*. Oneworld Publications.

Göritz, L., Stattkus, D., Beinke, J. H., & Thomas, O. (2022). "To reduce bias, you must identify it first! Towards automated gender bias detection."

Honneth, A. (1996). *The Struggle for Recognition: The Moral Grammar of Social Conflicts*. Studies in Contemporary German Social Thought

Holroyd, J., & Saul, J. (2018). *Implicit Bias and Reform Efforts in Philosophy*, University of Arkansas Press

Kamm, F. M. (2007). *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford University Press

Kay, M., Matuszek, C., & Munson, S. A. (2015). "Unequal representation and gender stereotypes in image search results for occupations," ACM CHI Conference on Human Factors in Computing Systems

Khader, S. J. (2017). "Transnational feminism, nonideal theory, and 'other' women's power," *Feminist Philosophy Quarterly*, 3(1)

Kourany, J. A. (2010). *Philosophy of Science after Feminism*. Studies in Feminist Philosophy

Müller, V. C. (2020). "Ethics of artificial intelligence and robotics," *Stanford Encyclopedia of Philosophy*

Tatman, R. (2017). "Gender and dialect bias in YouTube's automatic captions," *Proceedings of the ACL Workshop on Ethics in Natural Language Processing*

Waelen, R. and Wieczorek, M. (2022). "The struggle for AI's recognition: Understanding the normative implications of gender bias in AI with Honneth's theory of recognition," *Philosophy & Technology*

Young, I.M. (2001). "Equality of whom? Social groups and judgements of injustice," *The Journal of Political Philosophy*, 9(1), 1-18