Epistemic Injustice in the Age of AI

Martina Sardelli[*]

University of St Andrews

Artificial Intelligence (AI) is revolutionising our practices of distributing and producing knowledge. Though promising, these technologies also harbour the potential for corruption - a rising problem in this domain is that of injustice committed against women in the epistemic sphere. In our social framework, being regarded as a credible knower has become synonymous with the potential for self-actualisation: the realisation of one's potential. As such, the gender bias perpetrated by some AI systems is harming women in this domain. Additionally, biased software is barring them from accessing hermeneutical resources relevant to the understanding of their lived experience. Though still in its infancy, the problem should be urgently addressed by conceptualising ways in which a fairer AI could be engineered. Egalitarian ideas, specifically focused on equality of opportunity, seem to be promising avenues for future research and thought.

1 Introduction

Artificial Intelligence (AI) is the ability of computers and machines to perform tasks emulating those undertaken by the human mind, e.g. perception and decision- making, among others.[1] The development of these technologies in recent years has made their use intrinsic to the fabric of our daily lives, and AIs are now important components of our search engines, medical diagnostic tools and surveillance technologies.[2][3][4] Though AI software continues to define technological progress, the outputs produced by these systems can also perpetuate bias.[5] This, coupled with the power dynamics underlying gender discrimination, historically sustained by a social, political and economic infrastructure, are at the core of AI's intersection with epistemic injustice.[6] Epistemic injustice itself can be understood as a series of practices whereby knowers are wronged qua knowers, as well as practices which distort or impede the understanding of a knower, doing so from within a framework of established epistemic practices and institutions.[7][8]

In this essay, I endeavour to show how gender bias is perpetuated by AI through the lens of epistemic injustice. I claim that gender bias in AI is a problem which needs to be urgently addressed as it hinders women's capacity for self-determination and their ability to be perceived as credible conveyors and possessors of knowledge. I will do so first by

[*] My name is Martina, I come from Luxembourg and I am a final-year student at the University of St Andrews, studying Biology and Philosophy. Later in the year I am set to start a master's degree in social science and the internet. As such, I am particularly interested in the integration of AI into mainstream working practices and its repercussions! I also make the meanest focaccia.

[1] IBM Cloud Learn Hub, "What is Artificial Intelligence," 2020, accessed April 20, 2021, https://www.ibm.com/cloud/learn/what-is-artificial-intelligence.

[2] Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (NYU Press, 2018), accessed May 24, 2022.

[3] Fei Jiang et al., "Artificial intelligence in healthcare: past, present and future," *Stroke and Vascular Neurology* 2, no. 4 (2017): 230–243.

[4] Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ed. Sorelle A. Friedler and Christo Wilson, vol. 81, Proceedings of Machine Learning Research (PMLR, 2018), 77–91.

[5] Tolga Bolukbasi et al., "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," in *Advances in Neural Information Processing Systems*, ed. D. Lee et al., vol. 29 (Curran Associates, Inc., 2016), 1.

[6] Miranda Fricker, *Epistemic Injustice: Power and the Ethics of Knowing* (Oxford University Press, 2007).

[7] Gaile Pohlhaus Jr., "The Routledge Handbook of Epistemic Injustice," chap. Varieties of Epistemic Injustice, ed. Ian James Kidd, Jose Medina, and Gaile Pohlhaus Jr. (Routledge), 13.

[8] Fricker, 1.

presenting instances of gender bias in different AI systems and their detrimental effects on women's ability to transmit credible knowledge. Then, I will explore Fricker's account of epistemic injustice and how AI factors into the equation. A counterargument for AI's involvement in epistemic injustice focuses on AI's accountability in the moral arena, dodging claims of it being a cause of epistemic injustice. To this, I rebut that AI does not carry out epistemic injustice directly, but rather bolsters a social mind-set where the notion and practices of treating women as non-credible epistemic agents is buttressed. Lastly, I will address what a "fair" AI could look like using Rawls' theory of justice as fairness and Binns' notions of egalitarianism as applied to machine systems.

2 Laying the Groundwork: Instance of Gender Bias in AI

Instances of gender bias have been documented extensively in the AI literature from domains as disparate as precision medicine and civil surveillance, with concrete repercussions for women[9][10]. In many cases, gender and sex are viewed as confounding factors or unimportant, and often aren't factored into the training data which goes on to establish the architecture of an algorithm, *e.g.* for AIs doing precision medicine, "sex" is often omitted in training datasets[11].[12] Recently, important instances of sexist biases have been revealed through the inaccuracy of facial recognition software in identifying women, especially black women[13]. False positives generated by these technologies (*i.e.* cases in which a person is misidentified by the system) and their increasing deployment in the context of the criminal justice system threaten to seriously undermine the civil liberties of those affected in profoundly unfair ways[14]. In addition to gender, protected categories such as race, gender, ethnicity and religion are also at risk of being discriminated against. Though the programming of an unbiased AI necessitates factoring all these groups in, this essay will specifically focus on AI's *negative* bias against women. As such, any detailed analysis of racial, class or religious bias in conjunction with AI is beyond its scope.

Understanding the deleterious effects of biased AI in the medical and civil realm is a pressing issue, however, I hold that natural language processing (NLP), content moderation systems and automated résumé filters are particularly interesting to elucidate how artificially intelligent systems are tied to epistemic injustice in the purview of gender. Applications of NLP range broadly from recognition of speech to machine translation.[15] Most often, the algorithms for these systems have been shown to be insensitive to the different nuances of spoken and written language between genders, to the detriment of women. An oft- cited example is Bolukbasi *et al.*'s (2016) paper on biased word embedding NLPs. Word embedding is employed to capture semantic relationships between words, which are represented as vectors in a geometric space. Words whose semantic meanings are similar will have vectors located close to each other in this space[16]. These relationships can be shown using machine- completed analogy puzzles, e.g. "man is to king as woman is to *x*", where, in this instance, "*x*" equals "queen"[17]. Training of word2vec (an embedding technology) on a corpus of Google News (w2vNEWS) texts has shown clear perpetuation of harmful societal gender stereotypes in multiple instances. Bolukbasi *et al.*, claim that the sexist analogies (e.g. "she: diva= he: superstar"[18]) generated by the software trained on this data, seem not only to reflect but amplify pre-existing biases of the dataset on which the algorithm was trained[19]. Another incident where an AI, trained by engineers at MIT to label people and objects in images, went rogue, calling women derogatory terms in reference to sex workers, springs to mind[20].

As well as the perpetuation of harmful sexist and misogynistic stereotypes, the systematic censorship of women

[9] Davide Cirillo et al., "Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare," *NPJ Digital Medicine* 3 (81 2020): 1–11.

[10] Buolamwini and Gebru.

[11] Cirillo

[12] Katyanna Quach, "MIT apologizes, permanently pulls offline huge dataset that taught AI systems to use racist, misogynistic slurs," 2020, accessed May 6, 2021, https://www.theregister.com/2020/07/01/mit_dataset_removed/.

[13] Buolamwini and Gebru

[14] Buolamwini and Gebru

[15] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (Pearson Education, 2021), 1439.

[16] Bolukbasi, 1.

[17] Bolukbasi, 1.

[18] Bolukbasi, 2.

[19] Bolukbasi, 2.

[20] Quach

through content moderation software is cause for concern.[21] Binns *et al.* found content moderation systems to be less sensitive to female-labelled test data, generating more false positives: female-authored expressions which human moderators in the study did not deem offensive were more likely to be unreasonably classified as such and censored by the machine system[22]. A similar, but more covert, censorship is carried out by AIs which screen résumés: those featuring female job applicants from all-women colleges and including words such as "women's" were systematically targeted by Amazon's automated filtering system to be rejected.[23] Thus, the common, problematic denominator tying these systems emerges: they harm and handicap women epistemically, *i.e.* as credible knowers. In these cases, women are the subject of clichés which directly impeach their ability to communicate knowledge. More indirectly they play on sexist stereotypes, historically culpable for sustaining the perception of women as less rational and/or less capable of imparting reliable knowledge than their male counterparts, as well as excluding them from having the same opportunities on these very grounds. As I will explore in the next section, uncovering these biases is a key first step to comprehend how AI fits into the wider framework of epistemic injustice.

3 The "Stereotype - Prejudice - AI - Epistemic Injustice" Pipeline

Before delving deeper into the insidious ways AI can perpetuate epistemic injustice, it is important to illustrate the concept of injustice through an epistemic lens as well as the power dynamics at play when epistemic exchanges go awry. To lay the groundwork, Fricker defines social power as a form of power exerted structurally or by an agent whereby either has the capacity to affect others' actions in virtue of a specific social situation[24]. Though having a practical dimension, it also embodies what Fricker terms an aspect of "imaginative social co-ordination"[25], that is, shared beliefs about what it means to belong to a certain social identity, *e.g.* what it means to be a "woman". A corollary of social power, repurposed to focus on shared conceptions of social identity, is identity power, which is power exerted in relation to the conceptions of identity borne of our collective social imagination[26]. Epistemic judgements are fundamentally based on credibility judgements: the final aim of an epistemic exchange is to deem whether the knowledge conveyed is reliable or not (*i.e.* non- credible knowledge). Crucially, Fricker points out that heuristics are important to arrive at a final judgement and often involve the use of stereotypes[27]. Though she adopts the term neutrally[28], throughout this essay I will only use the term in its negative sense, just as I will only use "bias" in its negative connotation. As such, I will take stereotypes to indicate unreliable "widely held associations between a given social group and one or more attributes"[29]. Examples of sexist stereotypes include women's temperaments (*e.g.* "women are hysterical" etc.) but also extend to a more physical/practical dimension, *e.g.* brain make-up ("men's brains are wired to be better at physics" etc.). These unreliable "empirical generalisations"[30] typically prey on groups of people belonging to protected categories; whilst this essay focuses on gender, race and class (among others) have also historically been targeted by stereotypes[31]. Heuristic use of stereotypes is what paves the way to negative identity prejudice — that is: a generalisation based on the belonging of an individual to a social group which is impervious to counterevidence "owing to an ethically bad affective investment"[32]

Results produced by biased word-embedding, content moderation and résumé filter systems seem to use similar heuristic shortcuts: producing results which prey on negative stereotypes of women, historically a centrefold of our collective social imagination. Through the stereotypical portrayal of women as "homemakers" and "divas", preventing them from accessing traditionally male-dominated jobs (such as software engineering) and censoring them on platforms of public speech, it would be fair to say AI is spreading negative identity prejudice. In the next section I also argue this

[21] Reuben Binns et al., "Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation," in *9th International Conference on Social Informatics*, ed. Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri (Springer International Publishing, 2017), 3.

[22] Binns 2017, 6.

[23] Jeffrey Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," 2018, accessed May 5, 2021, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

[24] Fricker, 14.

[25] Fricker, 15

[26] Fricker, 15.

[27] Fricker, 31.

[28] Fricker, 31

[29] Fricker, 31

[30] Fricker, 32.

[31] Fricker, 33.

[32] Fricker, 35.

negative identity prejudice deleteriously affects women's epistemic status as they become tokens of suspicion and distrust, preventing them from being treated as reliable and valid possessors and/or conveyors of knowledge. For example, women may be seen as less qualified for software engineering positions, despite the fact they have the same credentials as their male competitors. However, their bid to apply might be assigned lower credibility in virtue of the fact that they are women, owing to their 'lesser aptitude in the scientific domain' or 'their final aim to become homemakers' etc. Having established an initial link between artificially intelligent systems and how they tie into epistemic injustice, I will now expand on how the association of the two has damaging effects on women's potential for growth and self-determination.

4 The Harms of Epistemic Injustice

Thus far, we have elucidated Fricker's notions of identity power, stereotype, negative identity prejudice and examples of AI's involvement in these. However, defining these terms doesn't capture why or in what capacity AI harms women's epistemic standing in our social framework. In this section I aim to further explain why stereotypes and prejudice are detrimental to self-development and self-actualisation as well as illustrating how these occur in tandem with the results produced by AI systems. As mentioned in the previous section, stereotypes and prejudice both operate on a practical social plane as well as an imaginative one, both of which require levels of social coordination to persist. Though harbouring the potential for change, our social imagination also bears the marks of previous prejudices, which can insinuate themselves into our collective social practices, thus becoming systemic[33]. But why *is* prejudice bad for our social discourse and collective imagination? Recall that prejudice operates on stereotypes, which in turn are "empirical generalisations". I believe that these are at the core of the prejudice problem: they hinder the epistemic development of a given social group by failing to treat the people belonging to it as *individuals*, missing out on member-specific truths on account of applying stereotype-heuristic shortcuts when making credibility judgements (whereas those belonging to these groups are likely to present their own idiosyncrasies and skills)[34]. For example, whilst some women may be more emotional than others, generalising that all women are hysterical in virtue of belonging to the social category "woman" will cause their lived experience not to be taken seriously even in cases where they express upset or hurt. This has been known to occur in doctors' offices, for example, where women's pain is often dismissed as melodrama, producing flawed diagnoses.[35] Not only is this harmful to someone personally, but it also presents an obstacle to understanding the truth[36]. This was apparent in the examples I mentioned, whether with biased software denying women certain job opportunities or their speech being censored for appearing too "emotional" on online fora. Here, the exclusion of women from contributing knowledge into the public domain is extremely detrimental not only in virtue of their human dignity but it also impeaches on their freedom of speech. By being unable to voice their knowledge, the silence they are forced into thwarts their chances of *potentially* being considered a credible knower.

Drawing on the Kantian notion that freedom of speech is essential to the authority of reason, Fricker asserts that censorship deprives us of the *capacity* for reason: the instrument of the human mind that has historically differentiated us from animals and become synonymous, if not nearly identical, with the human ability to convey knowledge[37]. Imparting knowledge is also instrumental to set the boundaries and parameters through which our social discourse unfolds. Those who have the opportunity to communicate knowledge, and are attributed with due credibility when they do so, have the power to dictate what constitutes acceptable discourse. This is why the censorship of women by AI content-moderation systems is so harmful: it betrays an insensitivity and indifference to the nuances of the way women express themselves and sets "norms of acceptability"[38] which are consolidated through sexist standards. At its core then, conveying knowledge is a human right: the opportunity to contribute and add to the social discourse is loaded with meaning and human value. When women are wronged in this (epistemic) capacity through biased software, it could be argued that their value and "currency" as human beings is belittled. This has the potential to stunt one's personal development, what Fricker calls a "process of social construction"[39], cramping the avenues in which a given individual can grow personally. In a sense, this

[33] Fricker, 87.
[34] Fricker, 33.
[35] Ian James Kidd and Havi Carel, "Epistemic Injustice and Illness," *Journal of Applied Philosophy* 34, no. 2 (2016): 172–190.
[36] Fricker, 44.
[37] Fricker, 44.
[38] Binns 2017, 2.
[39] Fricker, 59

process is centred on one's ability to make sense of one's lived experience, as gaining an understanding of this is a crucial step on the road to acquire knowledge, but how does AI impact this? It does so by excluding women from accessing the tools of social interpretation needed to make sense of some of their lived experiences. The following section will tackle this in further depth.

5 The Hermeneutical Consequences of AI and Epistemic Injustice

Though most instances of epistemic injustice detailed in this essay could be classified as testimonial injustices (*i.e.* attributing a negative, systematic, ethically- culpable credibility deficit to someone[40]), I can also envision AI systems perpetrating what Fricker terms "hermeneutical injustice", the second sub- category of epistemic injustice where the individual affected lacks the resources of social interpretation needed to makes sense of a consequential portion of their social experience[41]. Though it hasn't been in the purview of this essay to illustrate a taxonomy of epistemic injustice, the hermeneutical sphere of the latter presents interesting repercussions when explored through the lens of AI.

Let's trace back to the example of automatic résumé filters: a woman applies for an engineering job at a tech company with some stellar credentials which set her up to be a great candidate for the position she has applied for. She submits her résumé, which gets screened by an automatic résumé filter. This may find multiple instances of the word "women's" on her submission. The machine's model will most likely have been trained on the company's previous hiring data and as the job she has applied to has a tradition of hiring men for the position, the AI is highly likely to reject her application on the grounds of her being a woman. The woman is unsuccessful in her application but doesn't know an AI filtered her résumé. At this point, she is likely to feel bewilderment at being rejected from the job she was qualified for.

If she isn't aware of the machine algorithm used to make this decision and isn't privy to or can't understand the biases which have been encoded into the system, she will struggle to comprehend why she was rejected. I believe that understanding how this AI works and knowing that she was rejected by one are interpretive resources she would need to make head or tail of what Fricker describes as "an experience which it is strongly in her interests to render intelligible"[42]. This has the markings of hermeneutical injustice (though I use a slightly broader definition of "hermeneutical" than Fricker's): the rejected job applicant will probably feel confused, vulnerable and unsure about the integrity of her credentials, and to a certain extent, her identity as a person. The exclusion of a social group from professions in the hermeneutical sphere (hermeneutical marginalisation), is one but many realms in which hermeneutical injustice can track an individual (if the injustice is consistent in multiple different social domains in addition to the hermeneutical, the injustice is said to be systematic[43]).

An important clarification to make is that hermeneutical injustice is not carried out by an *agent* but rather is a lacuna intrinsic to our hermeneutical resources, caused by identity prejudice in the hermeneutical domain[44]. As AI is developing at break-neck speed but is still relegated to the realm of tech specialists, understanding the subtle but insidious ways it affects women's lived experience as a social group is a hermeneutical resource which is not available to the wider public.

Additionally, the problem of opacity in machine learning (ML) algorithms threatens to make this a reality for specialists too – at times, the way ML systems solve problems is not wholly intelligible even to those who have programmed them (also known as "The Black Box Problem").[45] Worryingly, the examples of epistemic injustice perpetrated by AI that I have presented all exclude the individuals they affect ("women") from being considered rational and significant

[40] Fricker, 28.
[41] Fricker, 148.
[42] Fricker, 148.
[43] Fricker, 156.
[44] Fricker, 169.
[45] Carlos Zednik, "Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence," *Philosophy & Technology* 34 (2021): 3.

members of society through the determent of their contributions from generating social meaning[46]. As I have outlined, epistemic injustice is persistent across multiple domains which extend beyond the social: preventing someone from creating meaning will also prevent them from creating significance and value, thus, I agree with Fricker's claim that this type of injustice has the potential to stifle the growth of important aspects of one's personal identity[47].

Through the perception of knowledge as a social currency of sorts, we are effectively buying into the idea that, in virtue of embodying a certain identity, some individuals (and their input) are intrinsically more valuable than others.

6 Mountain Out of a Molehill

Though there is empirical backing showing that AI is gender-biased, things get a bit murky when we try to assess the gravitas with which we should regard the outputs generated by these systems. The idea that AI makes consequential decisions affecting the epistemic standing of women *through* epistemic injustice rests on the assumption that negative identity prejudice is at play when an AI produces a sexist result. However, if we refer back to Fricker's definition of negative identity prejudice, one of its necessary features is that it must be ethically bad, or at the very least that there must be an "ethically bad affective investment" at play[48]. This begs the obvious question of whether AI can be ethically culpable or even said to have a morally questionable affective investment, thus, the larger question becomes whether artificially intelligent systems can be regarded as agents responsible for carrying out epistemic injustice against women. If they cannot be held ethically culpable, then perhaps it would be misguided to claim that their outputs embody negative identity prejudice in the way Fricker would have envisioned it. As such, it would be incorrect to say that AI carries out epistemic injustice in accordance with the parameters by which we have defined it in this essay.

Whilst the topic of moral and agential responsibility of AI deserves to be expounded further, this essay is not purporting to claim that software is *generating* or *causing* epistemic injustice, rather, that it is *perpetuating* it. These systems nurture preexisting stereotypes present in our collective social imagination and repackage them to symbolise impartiality through the guise of objective machine-driven outputs for the tasks they are programmed to perform. Though the focus of this work has mainly been on the outputs generated by AIs, these are arguably not at the root of the problem, instead, the issue seems to lie within the operation of training and refining the processes used by the systems to reach these (problematic) outputs. Gender bias in AI isn't a direct cause of epistemic injustice, but rather an *enforcer* — it lays the groundwork for a culture where epistemic injustice is justifiable within the social infrastructure we live in. The epistemic "wrongdoing" itself is perpetrated by those who buy into this and embrace the stereotypes promoted by biased software. Thus, the focus of the central argument I have put forward does not come down to *culpability* but rather *influence*.

More often than not, human judgement will come between the output of the AI machine and executive action; the result produced by an AI will be assessed critically by specialists before crystallising into something concrete. Thus, it may appear as though we are placing too much importance on the results churned out by AIs, despite them not being the direct cause of epistemic injustice. However important identifying causes may be, I believe this line of thinking is unconducive to recognising the real problem: it presupposes causation of the injustice to be the most relevant arbiter of importance for AI's involvement in the sphere of epistemic injustice. Though causation is undoubtedly key, I believe influence – i.e. AI's ability to reinforce, spread sexist stereotypes and censor women on a massive scale and creating a fertile environment for epistemic injustice to thrive, to be just as bad, if not worse, in its ability to justify and promote the spread of injustice, as it does so in a more insidious manner. Predominantly only discernible by a select few with the epistemic resources to do so (*i.e.* AI specialists). Problematically, this promotion of injustice occurs under the pretence of impartiality. As such, having established that AI is a powerful influencer of credibility outcomes for women in the province of epistemic injustice, (with some unpalatable consequences) how do we move forward? In the next paragraph, I propose a good place to start would be to conceptualise what a "fair" AI could look like and what a good way to conceptualise it could be.

[46] Fricker, 153, 161.

[47] Fricker, 169.

[48] Fricker, 36.

7 Building a Fair AI: New Solutions for New Problems

Most of us have an intuitive notion of what fairness entails, nevertheless, setting parameters to program a "fair" AI has proven to be difficult and the subject of much debate. In this section, I aim to expand on Binns'[49] account of fairness for machine systems and tie it to Rawls' concept of justice as fairness[50] as both present a compelling case to examine how to curb the AI's perpetuation of epistemic injustice. Naturally, questions arise when the idea of "fairness" is broached: what should the focus of a fair AI be (*i.e. should it maximise benefits for most or minimise harms for most?* etc.)? And when a focus is established, how could our chosen metric be quantified and practically applied to AI? The second question isn't as pertinent to the focus of this paper, as such it will not be tackled here. To address the first, however, Binns makes a compelling argument by positing that egalitarian norms can elucidate how algorithms are "unfair"[51]. Egalitarianism is an interesting avenue to explore as its doctrine lends itself well to answering "what should the focus of a fair AI be?" in focusing on the question "the equality of what?". For example, in the case of content moderation — should the chance one has of being censored be equalised regardless of gender? Or should equalising outcomes of the censorship be our focus?

The open-ended nature of this "the equality of what?" creates a debate concerning the application of egalitarian mores in "different social contexts"[52] and whether our answer to this question should be tailored according to the domain an AI system is operating in. Rawls engages with this question in the second principle of his thesis of justice as fairness. He believes we all have the right to a basic set of liberties and that these should provide the greatest benefit they can to disadvantaged members of society[53]. In addition, these basic liberties should be enacted by fostering conditions of equality of opportunity[54]. In Rawls' view, the latter hinges on the notion that everyone should have the same educational and economic opportunities regardless of the social "category" they were born into, as this is arbitrary (woman/man, white/black, rich/poor etc.)[55]. These things considered, we can explore a practical implementation of Binns' and Rawls' ideas using automatic résumé filtering software. Modelling equality of opportunity into these systems could avoid sexist outputs, for example, where software penalises résumés on the basis of containing words such as "women's" and "women's chess captain"[56][57]. If AIs were designed with "equality of opportunity" as a guiding principle, the epistemic injustice which undermines women as possessors and conveyors of specialist engineering knowledge in this case could be corrected, at least partially, as it would allow women to access the same resources (in this instance, jobs) as their male counterparts, granting them the chance to contribute to social discourse in a way that is deemed meaningful. Similarly, if we gave women equal opportunity to men in the arena of self-expression (without AI-mediated censorship), we could encourage a milieu receptive to the ways in which women communicate — enabling them to impart knowledge in a meaningful way. Though Fricker's theory of epistemic injustice is built on attributive (*i.e.* with the attribution of credibility) rather than distributive lines, I can envision "equality of opportunity" to be a good metric guiding attribution of justice on a case to case basis. After all, the adjustments which will need to be made to an algorithm's output will differ on the basis of other protected characteristics such as class, race or creed. However, knowing that our end goal is to guarantee that all women to have the opportunity to receive the credibility they deserve *qua* knowers will be central to stemming the perpetuation of epistemic injustice via AI.

Though this may not be perfectly aligned to what Rawls envisioned, as "equality of opportunity" predominantly concerns itself with economic goods, I still believe an interesting connection could be made between his second principle of justice as fairness and Fricker's conception of an "economy of credibility" and an "economy of collective hermeneutical resources". These terms are defined loosely in her work; however, if we envision knowledge as a currency of sorts (to accrue social status, power etc.), then a "credibility" economy or one based on "collective hermeneutical resources" will concern itself with the production, elaboration and consumption of knowledge. It seems to me that when women strive

[49] Reuben Binns, "Fairness in Machine Learning: Lessons from Political Philosophy," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ed. Sorelle A. Friedler and Christo Wilson, vol. 81, Proceedings of Machine Learning Research (PMLR, 2018), 149–159.

[50] Leif Wenar, "John Rawls," in *The Stanford Encyclopedia of Philosophy*, Summer 2021, ed. Edward N. Zalta (Metaphysics Research Lab, Stanford University, 2021).

[51] Binns 2018, 6.

[52] Binns 2018, 6.

[53] Wenar.

[54] Wenar.

[55] Wenar.

[56] Wenar.

[57] Dastin.

to be regarded credible *qua* knowers or achieve a full understanding of their lived experience they are effectively pursuing an epistemic currency of sorts, which allows their word to have due influence in the collective social discourse. Though coding these principles into an AI might prove to be complex – I believe it could be a viable solution to address the crux of the problem. AI is discriminatory and does not endow women with equal opportunity to create social meaning, doing so through faulty attributions of credibility and lack of epistemic resources available to them. In a competitive credibility economy, women must be able to fairly compete for the same resources as their male counterparts, and I believe a good place to begin would be by giving them equal opportunity to do so.

8 Conclusion

The problem of gender bias in artificially intelligent systems is rapidly gaining recognition. The increasing weaving of these technologies in our social and economic fabric has made it clear that further research of this phenomenon is warranted to address it and its ramifications. This essay proposed one such ramification to be epistemic injustice. I posited that gender-biased AI plays a role in perpetuating it by differentially censoring women on public platforms, sustaining sexist stereotypes which harm their credibility as knowers and preventing them from accessing the same opportunities as men on the basis of biased credibility judgements. Based on Fricker's case for epistemic injustice I have endeavoured to show that being able to impart knowledge is crucial for a person's self-development and that through their biased outputs, machine systems play a role in preventing women from creating valuable social meaning. This could be seen as an inflated view as, after all, AI cannot be said to commit an epistemic injustice as (per Fricker's definition) there is a debate surrounding its moral accountability. However, though AI may not be directly causing epistemic injustice it is creating an environment where it appears permissible and even normalised to do so, which, as I have argued, is highly problematic. In conclusion, I believe that to move forward we must build fairer AIs. I have argued that through Binns' and Rawls' ideas this would entail using an egalitarian framework to encode equality of opportunity into machine systems. And hopefully, loosen the chokehold of sexist stereotypes in our daily practices of receiving and producing knowledge.

References

Binns, Reuben. "Fairness in Machine Learning: Lessons from Political Philosophy." In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, edited by Sorelle A. Friedler and Christo Wilson, 81:149–159. Proceedings of Machine Learning Research. PMLR, 2018.

Binns, Reuben, Michael Veale, Max Van Kleek, and Nigel Shadbolt. "Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation." In *9th International Conference on Social Informatics*, edited by Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri, 405–415. Springer International Publishing, 2017.

Bolukbasi, Tolga, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings." In *Advances in Neural Information Processing Systems*, edited by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, 29:1–9. Curran Associates, Inc., 2016.

Buolamwini, Joy, and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, edited by Sorelle A. Friedler and Christo Wilson, 81:77–91. Proceedings of Machine Learning Research. PMLR, 2018.

Cirillo, Davide, Silvina Catuara-Solarz, Czuee Morey, Emre Guney, Laia Subirats, Simona Mellino, Anna Azzurra Gigante, et al. "Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare." *NPJ Digital Medicine* 3 (81 2020): 1–11.

Dastin, Jeffrey. "Amazon scraps secret AI recruiting tool that showed bias against women," 2018. Accessed May 5, 2021. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

Fricker, Miranda. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press, 2007.

Hub, IBM Cloud Learn. "What is Artificial Intelligence," 2020. Accessed April 20, 2021. https://www.ibm.com/cloud/learn/what-is-artificial-intelligence.

Jiang, Fei, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. "Artificial intelligence in healthcare: past, present and future." *Stroke and Vascular Neurology* 2, no. 4 (2017): 230–243.

Kidd, Ian James, and Havi Carel. "Epistemic Injustice and Illness." *Journal of Applied Philosophy* 34, no. 2 (2016): 172–190.

Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018. Accessed May 24, 2022.

Pohlhaus Jr., Gaile. "The Routledge Handbook of Epistemic Injustice." Chap. Varieties of Epistemic Injustice, edited by Ian James Kidd, Jose Medina, and Gaile Pohlhaus Jr., 13–26. Routledge.

Quach, Katyanna. "MIT apologizes, permanently pulls offline huge dataset that taught AI systems to use racist, misogynistic slurs," 2020. Accessed May 6, 2021. https://www.theregister.com/2020/07/01/mit_dataset_removed/.

Russell, Stuart, and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 2021.

Wenar, Leif. "John Rawls." In *The Stanford Encyclopedia of Philosophy*, Summer 2021, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University, 2021.

Zednik, Carlos. "Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence." *Philosophy & Technology* 34 (2021): 265–288.