CONFORMISM, IGNORANCE & INJUSTICE
AI as a TOOL of EPISTEMIC OPPRESSION[1]

ABSTRACT

From music recommendation to assessment of asylum applications, machine-learning algorithms play a fundamental role in our lives. Naturally, the rise of AI implementation strategies has brought to public attention the ethical risks involved. However, the dominant anti-discrimination discourse, often preoccupied with identifying particular instances of harmful AIs, has yet to bring clearly into focus the more structural roots of AI-based injustice. This paper addresses the problem of AI-based injustice from a distinctively epistemic angle. More precisely, I argue that the injustice generated by the implementation of AI machines in our societies is, in some paradigmatic cases, also a form of epistemic injustice. With a particular focus on AIs employed as gatekeepers of our epistemic resources, this paper shows how their epistemically conformist behaviour is responsible for the marginalisation and the ostracism of minorities' perspectives. Because it clarifies key structural flaws and weaknesses of current AI design, this paper helps make headway in critical discussion of current AI technologies. Because it forges new theoretical tools to understand forms of epistemic oppression, this paper also contributes to the advancement of feminist theorisation.

keywords: machine-learning AI, AI ethics, epistemic injustice, epistemic conformism.

word count 10,000

[1] I wish to thank [REDACTED] for helpful comments on earlier drafts of this paper. I also wish to thank an anonymous reviewer from *Episteme*, whose kind comments have greatly improved the quality of this paper.

I. AI, JUSTICE AND THE FUTURE OF RESEARCH

Consider the following examples:

> GOOGLE SEARCH Going through puberty, Irina has been experiencing new feelings for other girls her age. She turns to Google to try to understand more about these new feelings but what she finds is violent, over sexualised, cis- and heteronormative content. As a result, not only does she not find answers to her questions, but the research also instils in her a view of her sexuality that she doesn't feel is reflected as her own.

> ASYLUM SEEKER Negasi, a young Black man migrating from Ethiopia by way of Sudan, Chad and Libya, is seeking asylum in Germany. Their asylum application is processed via a new fully automated procedure just implemented by the Home Office. Despite having all the right credentials, and despite their story being true, Negasi's asylum application is unjustly rejected.

The widespread implementation of machine learning algorithms in services we rely on in everyday life has heightened the concern about new automated forms of oppression. The examples above show just a few paradigmatic cases of AI-based injustices systematically affecting members of minority groups. But the list is much longer. In *Algorithms of Oppression*, Sofia Noble gives a detailed analysis of the wide-ranging forms of sexist and racist prejudices that have been consistently found by typing racialised qualifications of individuals on the Google Search engine. More recently, translations from Hungarian, Finnish, Filipino and other gender neutral languages into English have revealed that Google Translate automatically

assigns female and male pronouns to genderless sentences according to stereotypical characterisations of genders. Translated to English, gender neutral sentences in Hungarian would read as follows: "She is beautiful. He is clever. He makes a lot of money. She bakes a cake. She is a cleaner. He is a professor. She is raising a child. She cooks. He is researching. He owns a business." (Ullmann 2021). But Google is not the only culprit. A study conducted by UC Berkeley on the algorithms employed to calculate targeted interest rates has found that information about borrowers (their geographical location, sexual orientation, spending habits etc.) allows the algorithms to profile ethnic minorities (who share comparable life conditions, such as living in financially isolated areas or being unable to do comparison shopping) and charge higher interest rates compared to White borrowers with comparable credit scores (Miller 2020). In criminal law, an investigation conducted in Florida by ProPublica (Angwin et al. 2016) on the scores assigned by AIs to rate a defendant's risk of future crime, has revealed that the machine was particularly likely to falsely flag Black defendants as future criminals, wrongly labelling them at almost twice the rate as White defendants, as well as mislabelling White defendants as low risk more often than Black defendants.

These cases display situations where AIs failed to function as they should. Google Search failed to provide adequate results for the search query inputted, the algorithms employed to calculate targeted interest rates failed to assess the creditworthiness of their applicants, and the risk scores have been found to unjustly favour White over Black defendants. These failures exacerbate unwarranted and unjust disparities, and generate harm. A working single mother that is denied a loan, for instance, is harmed financially, whereas Google's identity prejudices are liable to cause psychological or social harm.

Paradigmatic cases of AI-based injustice of this sort are now attracting the attention of the public and of the academic world, and have long been at centre stage for tech developers and researchers on the ethics of AI. Bracketing reactions of scepticism, the relevance of these cases is often taken to lie in

the challenges that they present to the fast-growing practices of development and application of AI-based technologies in our everyday lives.

These challenges have contributed to shaping an understanding of AI not only as a useful tool that we can rely on, but also as a culturally and historically determined product that we must learn to use responsibly. Indeed, concerns about ethics and social justice have always accompanied the history of technological advancement. Today, our culturally specific image of AI (Cave & Dihal 2020), its intrinsic biases (Noble 2018), and its connection with discrimination and harm (Bender 2021, Gandy 1998 and Adam 1998), are widely recognised to have a critical impact on our societies. These themes are now at the forefront of research on the ethics of AI, and constitute the theoretical premise of future development. The idea is that only by reflecting on the risks involved in its use can we hope to develop a more responsible relationship with AI in a way that can help us confront issues of social inequality, discrimination and oppression rather than exacerbate them.

Still, for the most part, critical theorising within AI has leveraged on a narrow and potentially damaging toolset, such as focus on singular 'bad actors' (Hoffman 2019). Take for instance the case of the report on the biases of AI proposed by Collett and Dillon in 2019. The report highlights concrete cases of gender prejudices in contemporary AI technologies. One of the cases discussed is that of automatic web-assistants, which, it has been found, are often characterised with stereotypical female attributes. The report proposes an informed and lucid analysis of the dangers associated with these kinds of practices, broadly connected with the perpetuating of stereotypical gender roles. In response, the possibility of overcoming this problem is envisaged by suggesting practical solutions (i.e., changing the gendered attributes of the assistant) and encouraging collaboration between AI developers and gender theorists.

Examples of this sort show that the way in which specific AIs are designed and function must be scrutinised if we want to prevent them from inheriting the bias of their developers, and that this cannot be done without a tighter interdisciplinary

collaboration. Indeed, the problem does sometimes boil down to identifying tech designers' and engineers' *dead spots* —that is, the unquestioned set of assumptions that is part of their cultural background (Snow 2018). But developer bias cannot be the sole cause of AI-based injustice. Oftentimes developers themselves fail to understand exactly why AIs develop certain prejudices. In these cases, there is a lack of interpretability of "black box" machine learning models —i.e., extremely long and complex sequences of algorithms whose functioning is impossible to predict for humans— that is not imputable to developer bias alone.

More in general, however, attention to developer bias has been criticised because it risks blurring our perception of the *structural* nature of the injustice at play —that is, both its connection with broader systems of oppression and in the sense in which AI-based injustice is necessitated by the very structure of AI systems in general. Contrary to this trend, an important strand of critical theorising within AI promotes a more systematic approach to AI-based injustice, interested in the multifaceted ways in which we interact with AI and actively contribute to strengthening and validating existing discriminatory social structures. As part of this 'structural turn' (Bagenstos 2006), work has been conducted to understand the limitations of narrow and mechanistic approaches to AI injustice (Hoffman 2019) and the importance of psychological (Krieger 1995) or cultural studies (Browne 2015) in giving central stage to broader concerns of social justice.

In line with this structural turn, I address the problem of AI-based injustice from a distinctively epistemic angle. More precisely, I argue that the injustice generated by the implementation of AI machines in our societies is, in some paradigmatic cases, also a form of *epistemic* injustice —namely, affecting us in our role as epistemic agents. The following discussion develops in three steps. First (§2), by looking at machine learning-based AIs employed as a gateway to our epistemic resources, I identify two interlocking concerns (i.e., what I call *toxicity* and *deficiency*) about their functioning and the training practices. These concerns, I show, stem from the

adoption of a fundamentally flawed principle of *epistemic conformism* in the very design of machine-learning based AIs. §3 leaves discussions about AI design behind to focus more specifically on the epistemic harms arising from their implementation and the way in which they contribute to reinforce structural oppression. More precisely, I argue that machine learning-based AIs erect barriers against AI-users, specifically targeting members of minority groups in their capacity as epistemic agents. In particular, following Mason (2011), I show how, seen as a form of 'hermeneutical lacuna', the toxic deficiency of AI harms agents as knowledge *seekers*, while understood as a form of 'white ignorance' (Spivak 1999, Mills 1997, Martín 2021) it risks harming them as knowledge *givers*.

Here, the importance of the discussion for feminist theorisation is brought to light as two new forms of epistemic injustice are individuated: what I call *zetetic injustice* and *testimonial spurning*. The former, an expansion on Fricker's (2007) taxonomy, concerns agents who are unjustly obstructed in their attempt to carry out meaningful inquiry. The latter, building upon Kristy Dotson's (2011) notion of epistemic violence, and akin to her notion of testimonial quieting, concerns agents who are unjustly prevented from obtaining what it is in their right to obtain with their words.

II. BIASSED DATA AND EPISTEMIC CONFORMISM

The quantity of content produced and stored online is vast. According to rough estimates, it amounts to over 30 zettabytes. To give an idea of the size of this, consider that streaming it using the fastest networks available would take over 2000 years[2]. The exponential increase, over the last few decades, of online data has urged experts to come up with solutions to help us navigate it comfortably. This challenge has been met by making recourse to intelligent 'sorting machines', trained to recognise and group together recurrent patterns of information among

[2] Statista Research Department (2022) https://www.statista.com/statistics/871513/worldwide-data-created/

vast pools of data. Today, most of the streaming services (Instagram, Netflix, Spotify, Youtube), systems of recommendation (Google, Baidu) and rating services (credit and assurance risk assessment, medical and legal services, etc.) that we use everyday are underpinned by the functioning of these machines, specifically designed to supervise and mediate access to specific epistemic environments —i.e., pools of online data. The rise of AIs of this sort has been possible thanks to the introduction, in the early 90s, of a sophisticated method of data analysis known as *machine learning* (ML). Machine learning is a term used to refer to a technique that consists in applying long strings of algorithms —long sequences of functions, or rules, that extract predictions from a given set of input values— and statistical analysis to numerical input values to produce numerical or binary (yes/no) outputs. More broadly, the term "machine learning-based AI" is generally used to refer to long strings of complex functions that have information (e.g., a search query) as input and output predictions (Hao 2018). ML-based AIs are thus essentially *predictive* machines. On the input of our online interactions (clicks and likes) and personal information (geographical location, gender, age, occupation etc.), ML algorithms extract and use patterns to predict our future clicks and likes.

To see how this works more precisely, consider the case of Spotify. Spotify is a music streaming service equipped with a content recommendation system powered by ML algorithms. When you listen to a song (album, artist or podcast), the system compares information about that song (e.g., the artist, producer, etiquette, genre, rhythm, melody, pitch) with patterns of information about content in Spotify's database that share similar characteristics. In this way, after we listen to a song by the Beatles, it may suggest songs by John Lennon (in virtue of the similarity between the song's artist and artist suggested) or by The Kinks (in virtue of a similarity between their pitch and melodies) and so on. Spotify's functioning depends on the fact that the machine has been trained to recognise the similarities between the input information (the question we ask Google, the song we listen to on Spotify, the series we watch on Netflix, the

digital request we submit for a loan etc.) and the trends and patterns of information present in their database ("hip hop music", "philosophy podcast", "drama series" etc.).

Patterns and trends are thus crucial to Spotify's ability to read and interpret the input information and output the prediction. In modern ML-based AIs, patterns are individuated through a procedure known as *data mining*, which consists in the sorting of information through a process of statistical analysis. Statistical sorting is a crucial part of ML-based AI functioning, as it provides a rationale (i.e., *statistical frequency*) for the identification of the trends and patterns that are then used to read and interpret the input information and finally output the prediction. AI's reliance on statistical analysis makes another factor crucial for its well-functioning, namely the *size* of the training data. Data is the raw material that is fed into ML-based AIs and that fuels its sorting engines. Because these machines function by selecting and identifying data patterns on the basis of their statistical frequency, the ability to identify diverse and reliable trends depends on the availability of large pools of data. The bigger the pool, the more solid and varied the trends available, and so the more accurate and adequate the machine's predictions.

To summarise, then, the more statistically robust a piece of information —i.e., the larger the amount of information that bears a relationship of close similarity with it— the more likely it is that the machine will be able to read and understand that piece of information (i.e., a search query), and provide accurate responses to it in the future. To simplify things, we can call the relevance a piece of information has with respect to the the machine's epistemic and hermeneutic abilities (i.e., the ability to read, understand and respond to that input information) *epistemic relevance*, and say that the epistemic relevance of a piece of information is just a function of its statistical robustness. The more common the input information, the more likely it is that it shares similarities with patterns already identified by the AI and present in its epistemic environment, and so the greater the machine's ability to read and understand it and output predictions that are adequate and accurate.

In what follows, the focus of my discussion will be on ML-based AIs regulating access, participation and contribution to shared online resources. Sometimes, I will be interested in this role as consisting in mediating the retrieval of information from online pools of data. In this case, the discussion will focus on search engines and recommendation systems, like Google, Spotify and Youtube, whose role is to help users navigate resources stored online. Sometimes, I will be interested in ML-based AIs as regulating participation in epistemic environments and practices —like when, for instance, AIs are employed for the assessment and evaluation of the liability, creditworthiness, or credibility of their users. More in general, then, in this paper I will be looking at AIs as gatekeepers of particular pools of information within our broader epistemic environments[3].

In light of the increasing importance ML-based AIs are assuming in everyday life as gatekeepers of shared knowledge, ML-based AI's reliance on data mining and statistical sorting procedures has been the focus of harsh criticism[4]. In a recent article, Bender et al. (2021), refer to AI employed in the generation of text (like the recent GPT-3, BERT and Switch-C) as a *stochastic parrot*, on the grounds that machine functioning consists in "haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning" (2021, 617). It would be misleading, they warn us, to take AI intelligence to be based on the machine's ability to engage in genuine critical thinking, since all it boils down to is the mere parroting of the most common trends of information detected in its training data. Assuming that Bender is right, and that it is true that AIs do have features justifying the association between the hermeneutical abilities of

[3] Thinking of AI as gatekeepers of shared online epistemic environments does not exaggerate the importance of AI in our everyday lives. Considering that a great deal of information we possess today is stored online and accessed via AIs, their importance can hardly be overstated. Moreover, thinking of AIs as gatekeepers is not to think of AI as the sole gatekeepers of *all* epistemic resources.

[4] Bender et al (2021), Krieger (1995), Hoffman (2019), Gandy (1998)

ML-based AIs and stochastic parrots, I want to propose a characterisation of the functioning of ML-based AIs in analogy with conformist attitudes —i.e., as instantiating a tendency to value or endorse attitudes and behaviours that are commonly accepted *simply because* they are commonly accepted. More exactly, what I want to suggest is that ML-based AIs could be characterised as exhibiting something in the vicinity of the following feature[5]:

> *Epistemic Conformism* The tendency to only treat as epistemically relevant information that is statistically dominant *because* it is statistically dominant,

where the epistemic relevance of a piece of information is just a measure of the likelihood that that piece of information is understood and offered an adequate response by a ML-based AI. The thought here is that AIs' conceptual repertoire is based on the resources present in statistically dominant trends; by referring to AIs as epistemically conformist, then, my aim is to formalise the idea, implicit in the idea of AI as 'stochastic parrots', that AIs simply *mimic* common trends present in their training data[6].

But referring to AI's functioning as conformist, to the extent that it may suggest that AI machines *merely* mirror the content and structure of our linguistic practices and conceptual

[5] Clearly, I take the claim that AI machines do as a matter of fact possess this trait to be contentious as it depends, among other things, on the plausibility of treating AIs as epistemic agents. But this should not constitute an obstacle to the point I want to make here, which relies merely on the plausibility of recognising some degree of analogy between the functioning of ML-based AIs and conformist attitudes conceived along these lines.

[6] Note that, despite their similarity, the notions of 'stochastic parrot' and 'epistemic conformism' are importantly different. First, because the notion of 'stochastic parrot' is used by Bender to criticise the idea that ML-based AIs can be thought of as competent language users and that they can understand what they are saying. My notion of 'epistemic conformism', on the other hand, is neutral with respect to issues concerning whether ML-based AIs are competent language users, whether they understand what they are saying —or, for that matter, about the relationship between the two. With the notion of 'epistemic conformism', instead, I wanted to capture the distinctively epistemic principle underpinning the functioning of ML-base AIs. In this sense, and in line with the general scope of this paper, we could arguably say that 'epistemic conformism' could be taken to clarify the epistemic aspect of the notion of 'stochastic parrot'. (I wish thank an anonymous reviewer for bringing up this point)

resources, can be misleading. ML-based AIs are not neutral tools: they play an active role in shaping the resources to which they mediate access. Consider again the case of Spotify. For those who rely on Spotify as their main access to multimedia content, the Spotify recommendation system influences the distribution of the contents and their availability by singling out those patterns in one's listening preferences that bear closer similarity to the patterns that the system deems more relevant, and suggesting predictions based on that. What's more, because such relevance is measured in terms of statistical robustness, information will be distributed in such a way as to make more readily available 'trendy' information, while unpopular content will be more difficult to identify and retrieve. Think for instance of the different results you obtain depending on the kind of search query typed into the Google Search box. The accuracy and adequacy of search hits relating to common queries (e.g., "interpretation of the song 'Hey Jude', by The Beatles") are much higher and diversified than that of queries relating to a domain or a topic that doesn't get as many search hits (e.g., "interpretation of the song 'Gli Impermeabili', by Paolo Conte").

Because it measures epistemic relevance on the basis of statistical robustness, we can predict that ML-based AIs' epistemic conformist functioning will lead to the formation of knowledge-gaps and interpretative lacunae, affecting the machine's ability to read, understand and respond to minoritarian information (i.e., pieces of information that bear little to no similarity to statistically robust patterns). As a result of their conformist behaviour, ML-based AIs appear to manifest a fundamental lack of interpretative power —that is, a structural *deficiency*— with respect to minoritarian vocabularies, language norms and systems of meaning. Crucially, because it stems from its epistemic conformism, AI's deficiency is part of the machine's very *design*. It is the AI's conformist behaviour that, because it grounds the epistemic relevance of a piece of information on its popularity, encodes a fallacious epistemic principle leading to the systematic marginalisation of minoritarian information and the formation of lacunae in our

epistemic environment[7]. This principle underlies the functioning of the sorting mechanism whereby patterns of information are identified, and that in turn determine the machine's ability to provide adequate and accurate predictions. Being marginalised, patterns of minoritarian information fail to be identified, and thus fail to form part of the machine's interpretative tools, which is in this sense importantly *deficient*[8].

Notice at this point that all I've said so far tells us nothing about the normativity of the environment that is thus affected —that is, whether AI's conformism affects it for the better or for worse. In fact, conformist attitudes are in some respects *neutral*: although they do impact the distribution of information in a determinate environment, they do so on the basis of a sorting principle that doesn't take into account its quality. Indeed, AI's conformist behaviour might uphold *good* just as much as *bad* epistemic environments —the minoritarian views screened off by the algorithms may be climate scientists' opinions on climate change just as much as neo nazis' claims about national identity, and whether AIs' conformism ends up upholding either will depend on empirical facts about the epistemic quality of the statistically dominant strands of information.

A recent study conducted by Emily Bender and her team (Bender et al. 2021), focussing on Google's norms of implementation (although it refers to practices that are now widely standardised) has revealed that, ML-based AIs' need of large swathes of data is met by relying on the largest database available today —namely, online networks and communities such as Reddit and Wikipedia. In particular, the aim of Bender's

[7] Note that this is true even if conformist ML-based AIs do provide accurate responses in most cases. The problem, in fact, does not have to do with the overall rate of successful responses given by AIs, but with the badness of their conformist design itself, which causes the AI to make epistemically relevant distinctions between types of information on the basis of facts that should not matter *epistemically* —namely, their statistical frequency. I thank an anonymous reviewer for pressing me to clarify this point.

[8] Note at this point that the word 'minoritarian' here is used in a strictly statistical sense. The content that is marginalised is simply content that fails to meet the threshold required for it to be read and adequately interpreted by the algorithms. A connection between minoritarian content and content expressing the world-view of non-dominant groups is proposed towards the end of this section.

article is to highlight the dangers that are associated with such practices. These span from the environmental costs of the data mining procedures (linked to the extraordinary processing power they require) to the way AIs are perceived in our society (ML-based AIs capacity to analyse and produce intelligible pieces of text gives the false impression that the machine can understand natural language). More importantly, however, their work draws attention to a fundamental problem connected to the quality of the information that is gathered from these sources. These concerns primarily stem from the consideration that access to the internet and its use are a prerogative of people from richer countries, and is more substantial among the wealthy White male youth (Bender et al. 2021, Roser & Ritchie & Ospina-Ortiz 2015). "GPT-2's training data" they argue, "is sourced by scraping outbound links from Reddit, and Pew Internet Research's 2016 survey reveals 67% of Reddit users in the United States are men, and 64% between ages 18 and 29. Similarly, recent surveys of Wikipedians find that only 8.8–15% are women or girls" (Bender et al. 2021, 613). Moderation practices regulating access to subsamples of the internet are also cited in this study as having a substantial discriminatory impact. A research conducted using digital ethnography techniques on Twitter (Jones 2020), for instance, has shown that people on the receiving end of online discrimination are more likely to have their account suspended than those perpetrating it.

Epistemic environments where discriminatory, biassed and harmful contents and norms prevail are *toxic* epistemic environments. Since empirical research gives an image of our shared online resources as expressing the world-view of dominant groups, reflecting their biassed, harmful, and often colonising view of the world, our shared online resources are thereby *toxic* in this sense —i.e., in the sense that they are permeated with contents and norms of bad epistemic quality.

In summary, then, I've pointed out two main concerns regarding ML-based AI design and implementation practices strategies. Because of the corruption of online resources that are employed to train AI machines, the epistemic environments to which AIs

mediated access are often epistemically *toxic*; and because of the knowledge-gap generated by AI's conformist behaviour, such machines discriminate against trends of information that are statistically weaker, and is thus unjustly *deficient*. Note however how, although distinct, it is in combination with each other that toxicity and deficiency influence the implementation of ML-based AIs. In particular, in what follows I will be interested in the way in which toxically deficient AI are responsible for the epistemic marginalisation of the language norms and vocabulary of minority groups. How so? Consider again AI's deficiency. Because it measures epistemic relevance on the basis of statistical robustness, I argued, the epistemic conformism of ML-based AIs leads to minority voices being systematically marginalised —that is, it prevents them from contributing equally to the formation of the shared meanings, concepts and interpretative tropes that operate within society. Because the statistical weakness of online content expressing systems of meanings of minority groups is an empirical fact (as per the epistemic toxicity of AI), it is possible to see how, more often than not, the minoritarian voices that end up being marginalised due to AI's deficiency are precisely the voices of members of minority groups.

In the next sections I turn my attention from the design and function of AIs to issues arising from their implementation. In particular, the aim will be to identify the ways in which AI's toxic deficiency contributes to set up barriers to epistemic agents as knowledge *seekers* and knowledge *givers*.

III. HERMENEUTICAL LACUNAE AND WHITE IGNORANCE

Take again the two cases considered at the outset. In GOOGLE SEARCH Irina, a young girl who wants to learn more about her own sexuality, is not only unsuccessful at finding content that can help her understand her own sexual experience, but throughout her research she is also repeatedly exposed to violent and overly sexualised content. ASYLUM SEEKER, on the other hand, describes the case of an Ethiopian man, Negasi, whose

asylum request is rejected by a new fully automated system implemented by the German Home Office. In both cases something went wrong: Irina and Negasi's pursuit of their epistemic goals (i.e., to inquire into a topic or to acquire or transmit a piece of information) have been unjustly trumped by barriers set up by the technologies they have relied on to achieve them. Irina's inquiry was unsuccessful, and so was Negasi's application.

Crucially, these barriers have been erected by the toxic deficiency of ML-based AIs. Irina's search queries are interpreted in the light of the categories extracted from the toxic dominant trends which do not include the kind of statistically non-dominant information Irina is after. The same discriminatory content also constitutes the interpretative categories through which Negasi's application is evaluated and the grounds on which it is rejected, since ML-based AI assessed Negasi's testimony not only against its actual credential, but also as a function of prejudiced assumptions present in the training data —in this case, say, the prejudiced thought that Black people are more prone to deception, and thus less likely to give accurate testimony.

My goal in this section is to show that these two examples stand for two paradigmatic ways in which AI's toxic deficiency causes distinctive epistemic harms. I will point at two main ways in which this deficiency can affect the epistemic agency of the members of an epistemic community in harmful ways: as a hermeneutical lacuna, and as a form of active ignorance. Talking about AI's toxic deficiency as a *hermeneutical lacuna*, I will show how this deficiency sometimes impairs the ability of members of minority groups to inquire into a topic or obtain knowledge regarding matters that are meaningful to them, thus harming them as *knowledge seekers*. With its identification with a form of *active ignorance*, on the other hand, I will be interested in understanding the way in which AI's toxic deficiency is responsible for perpetrating epistemic violence against members of minority groups by interfering with their ability as *knowledge givers*. Each of these barriers, I argue, becomes the source of a

new form of epistemic harm. I call *zetetic injustice* the one resulting from barriers erected against epistemic agents as knowledge seekers, and *epistemic spurning* the one erected against epistemic agents as knowledge givers.

It is important to bear in mind, as the discussion goes on, that the aim of my argument is not to establish that the design and functioning of ML-based AI is detrimental to minority groups *exclusively*, nor that the harms I am concerned with here are the *only* AI-based harms we should look out for. The general scope of this part of the article is to advance feminist and critical race theorisation by showing some of the ways in which ML-based AIs risk contributing to worsening the oppression of minorities in society.

III.1 Hermeneutical Lacunae and Zetetic Injustice

Based on the proposed characterisation of toxically deficient AI, the most obvious sense in which AI appears to be deficient is arguably with respect to the conceptual resources required for understanding, interpreting and adequately predicting requests pertaining to minoritarian preferences and patterns. How so? ML-based AI manifests conformist behaviour, which consists in a tendency to treat statistically robust patterns of data as epistemically relevant precisely in virtue of the fact that they are statistically robust. As a result, statistically weaker patterns of information, which fail to meet a statistical threshold, are systematically screened-off, and thus prevented from contributing to shaping the machine's interpretative resources. Because of the toxicity of the data scraped off the internet and used to train the AI, moreover, statistically weaker patterns invariably end up corresponding to the meanings, norms and interpretative tropes of minority groups.

ML-based AIs, then, lack the necessary conceptual competence to understand and interpret inputs from minority groups. If this is true, we should expect that attempts made from members of minority groups to access information that is relevant for them through ML-based AIs will fail systematically. In fact, this is precisely what goes on in GOOGLE SEARCH —because of the epistemically conformist behaviour displayed

by Google's algorithms, which tends to read and interpret input information in the light of the categories extracted from the dominant trends, the overwhelming majority of information Irina gets access to concerns the heteronormative and often violent forms of sexual expression that are most common among the majority of Google users, and that aren't helpful to her to make sense of her own sexual experience. In other words, the bias ingrained in the functioning of the Google Search engine prevents Irina from obtaining information that is relevant for her to understand aspects of her own identity.

Put this way, the case will strike those who are familiar with Miranda Fricker's notion of epistemic injustice as bearing close similarity to her characterisation of *hermeneutical* injustice. According to Fricker (2007), hermeneutical injustice is a particular form of injustice suffered by one as an epistemic agent, concerning one's ability to access meaningful information. More exactly, Fricker takes hermeneutical injustice to occur when prejudice ingrained in the body of shared interpretative resources hinders one's ability to obtain knowledge that is necessary to express oneself and to be understood. The prejudice is manifested in the form of gaps, or lacunae, in our hermeneutical resources —i.e., the tools, such as concepts or tropes, we use to make sense of our own experience. Now, since these lacunae occur at the level of our shared resources, and are formed and sustained by our collective meaning-making activities, hermeneutical injustice often concerns structural features of our communicative exchanges and social practices. The hermeneutical marginalisation of women, for example, is typically invoked to explain the lack, until very recently, of a fully-formed, shared concept of sexual harassment in our collective hermeneutical resources. Fricker's though is that, prior to its acquisition, victims of sexual harassment didn't have the conceptual resources required to come to know a fundamental part of their experience, and so to make sense of it.

Similarly, it seems plausible to describe GOOGLE SEARCH as a case where Irina is prevented from obtaining knowledge that is important for her to make sense of her own experience. Crucially, she is thus obstructed by a lacuna in the

shared online hermeneutical resources, a lacuna that is due to the predominantly discriminatory language and biassed world-views ingrained in the data used to train ML-based AIs like the Google Search engine[9]. If this is correct, we can derive an important conclusion from this argument. That is: because the hermeneutical lacuna present in our shared online resources is a direct consequence of the very functioning and training practices of ML-based AIs, epistemic injustices of a hermeneutical kind like the one suffered by Irina, are not just unlucky byproducts of developers' biases, but a *systematic feature of the design of AI design*.

A closer look at this case, however, seems to suggest another sense in which Irina is harmed in their capacity as a knower. First of all, notice that the prejudice ingrained in the machine's interpretative resources doesn't just prevent Irina from *obtaining* the valuable piece of information she's after. Recall how, in her attempt to find out more about her own experience, Irina not only fails to find what she's looking for, but her very attempt to *search* for it is repeatedly frustrated. Her queries, concerning vocabulary and concepts that aren't recorded in statistically robust trends, are systematically redirected to mainstream ones, often exposing her to violent heteronormative contents. On the face of it, then, it looks as though the hermeneutical injustice suffered by Irina is only the backhand of another barrier set up by the Google AI, this one against her attempt to conduct meaningful inquiry. To see better the kind of harm at play here consider the following case.

[9] I believe that a criticism moved by Rebecca Mason (2011) concerning the limits of Fricker's model applies here. In a nutshell, this criticism is that "[a] gap in dominant hermeneutical resources with respect to one's social experiences does not necessitate a corresponding gap in nondominant hermeneutical resources." (2011, 300). Mason's point is even more obviously true in cases like GOOGLE SEARCH, where the pool of shared resources is the even more restricted pool of online resources. While I agree with Mason, I think it is important to add how, even in the light of this consideration, it is still hard to overestimate the importance of dominant pools of information in one's epistemic life. This is largely because dominant knowledge is often also *sanctioned* knowledge, and is thereby granted special epistemic status. This I think is an important reason why the point made by Fricker retains special relevance even if, as Mason rightly points out (echoing Mills), hermeneutical resources are often already available outside through non-dominant channels.

> SWEETGRASS Laure is a final year botany student, and she needs to find a topic for her dissertation. She has long been interested in indigenous harvesting practices, and over the years has collected various testimonies from indigenous experts regarding techniques of harvesting that, they say, would preserve and improve the quality of sweetgrass crops. She finds that experts are polarised on the topic —some say crops benefit from a harvesting technique involving the cutting of sweetgrass stems near the roots, while others favour the method of uprooting. Finally, she makes up her mind and decides to dedicate her thesis to settling this disagreement. When she presents her idea to the school, however, the academic committee refuses her research proposal on the grounds that, they say, it goes against the known scientific fact that harvesting *damages* crops. The committee also undermines the validity of the testimony of the experts gathered by Laure, on account of the fact that they are mostly old indigenous sweetgrass pickers and basket-makers, not scientists, and encourages her to focus her thesis on another project. As it turns out, research conducted several years later reveals that the scientific consensus is wrong and that, for some plant specimens like sweetgrass (like Laure had thought, backed by the knowledge of expert indigenous sweetgrass pickers) some types of harvesting *do* improve the quality and quantity of the crop[10].

Laure has evidence, gathered through years spent with people in communities in close contact with sweetgrass, suggesting a promising line of inquiry. Yet this evidence is present only in small centres at the periphery of the main streams of knowledge production and diffusion. Members of the academic committee, as gatekeepers of the mainstream, reject Laure's proposal on the

[10] This case is taken from Robin Wall Kimmerer's 'Braiding Sweetgrass' (2013)

ground of a conformist decision —i.e., the decision to consider as scientifically relevant and worthy of pursuit only research that complies with mainstream assumptions and knowledge. Despite promising, Laure's inquiry is thus unjustly frustrated.

Like Irina's, Laure's inquiry attempt is also threatened to be undermined or absorbed into more mainstream patterns. Like Laure's, Irina's attempt to conduct research is also unjustly frustrated by the conformist resolutions of the gatekeeping authorities. While the gatekeeping role in Laure's case is played by the scientific committee, in Irina's that role is occupied by Google algorithms. In both cases the academic committee and Google algorithms are equally responsible for perpetrating the same form of injustice: by getting in the way of Irina's and Laure's inquiry and obstructing exercise of their epistemic autonomy, they are responsible for harming the two women in their capacity as knowers. More precisely, because it concerns their distinctive ability to conduct meaningful research, question and, more generally, inquire into matters that are relevant for them, I propose to call this particular form of wronging *zetetic injustice*.

The concept of zetetic injustice I have in mind falls under the broad category of epistemic injustice (although perhaps not in the sense this is used by Fricker). For example, I take that, thus characterised, zetetic injustices can be taken as just another variety of epistemic injustices, standing side by side with testimonial and hermeneutical injustices[11]. Like other forms of epistemic injustice, zetetic injustice also concerns one's epistemic conduct, and it too has identity prejudice as a key ingredient —although the examples discussed seem to suggest a pretty loose characterisation of prejudice as something that has less to do with one's cognitive commitment, as Fricker thinks, and more with structural flaws of one's epistemic environment.

On the other hand, zetetic and epistemic injustices naturally differ in important respects —most saliently, regarding the fact that zetetic injustice does not concern the obstruction

[11] I recognise that this may be contentious, as the relationship between the epistemic and the zetetic is a matter of ongoing debate. However, I don't think that anything substantial about my position here relies on this commitment.

of knowledge transmission or acquisition. In SWEETGRASS, the barrier put up by the academic committee against Laure's proposal does prevent the acquisition of valuable knowledge —the knowledge that, at least for some plant specimens, harvesting can improve the quality of the crop. The zetetic wrong Laure is a victim of, however, doesn't depend on that. She would have been wronged in her capacity as an inquirer even if subsequent research confirmed the scientific consensus, or if it proved uninformative. What matters for the kind of injustice at play, instead, is merely that Laure ends up being obstructed in her attempt to carry out the research itself[12]. The (implicit or explicit) barriers raised against an inquirer will vary depending on the context, but will typically function to mislead or misdirect the investigation. In SWEETGRASS, for example, the obstruction is caused by the faulty functioning of conformist and sectarian academic practices, and involves things like discouraging the researcher from carrying out her research, offering alternative research opportunities, possibly cutting her funding and so on. In GOOGLE SEARCH, the obstruction (caused by the problematic functioning of the Google Search algorithm I have described, such as the toxic deficiency and the conformist mechanisms that systematically produce it) involved offering inadequate responses to the search queries, providing misleading information, and attempting to reconduct the investigation towards more mainstream topics.

In summary, then, looking at the epistemic impact of ML-based AI reveals that the structural faults of the machine's design lead to the systematic production of particular forms of injustice. More precisely, it looks as though the hermeneutical lacunae in our shared online resources, due to AI conformist behaviour, are susceptible to cause those who rely on them to suffer from injustice of *hermeneutical* and *zetetic* varieties. Because they concern members of minority groups' failure to obtain resources that are meaningful for them, or even to inquire into them, I take these injustices to broadly consist of impairments they suffer as knowledge seekers.

[12] Naturally, the inquiry must also respect some basic zetetic norms —like, say, that one ought not to set out to inquire into whether X if one already knows that X.

III.2 White Ignorance and Epistemic Spurning

Because it is due to the toxic deficiency of AI design, the presence of hermeneutical lacunae, I have argued, tends to epistemically harm, for the most part, members of minority groups. However, it would be a mistake to think that minority groups are thereby relegated to a position of epistemic inferiority. This point, raised for the first time explicitly by Du Bois (1989 [1903]), and picked up and articulated more recently by Charles Mills (1998), reflects the fundamental insight of standpoint epistemology that "social privilege does not necessarily entail epistemic privilege" (Mason 2011, 301). In fact, the opposite is often and in crucial respects true: occupying a position of social disadvantage often puts one in a position of epistemic privilege. One influential way of explaining how this is the case is in terms of Charles Mills' notion of 'Racial Contract' (1998). According to Mills, dominant groups tend to think of their social organisation in terms of ideal, fundamentally *just* systems of meaning that exclude the possibility of the existence, from their very inception, of forms of oppression, injustice and discrimination. For this reason, an epistemic *asymmetry* is created between dominant and oppressed groups, whereby the former group, because these gaps and shortcomings are constitutive of their own world-view, tend to systematically fail to understand or (literally, according to Mills) perceive them. The latter group, instead, who often end up suffering from the lacunae in the fabric of the shared epistemic resources, and in virtue of the harm they often encounter (although not necessarily because of it, or not exclusively) become aware of them[13].

If true, Mills framework can offer a powerful interpretative key to the case. Recall that our online resources are constituted, for the most part, by content expressing the biassed, often discriminatory language and norms of wealthy White men.

[13] Note that this is not to say that, simply by virtue of being a member of a minority, one automatically obtains this kind of awareness, nor that all instances of injustice are revelatory of structural oppression. Yet, because they are oppressed, members of minority groups are in a position of natural advantage when it comes to obtaining awareness of the injustices and lacunae of dominant systems of meaning —as per the key insight of standpoint epistemology discussed earlier.

The toxic deficiency inherent in their own world-view, then, becomes manifest to members of non-dominant groups as a consequence of the (hermeneutical and zetetic, for instance) injustices they suffer, and which are caused by the knowledge-gaps and lacunae in the shared online hermeneutical resources.

Crucially, however, despite the new awareness acquired, because of the very design of ML-based AIs —which are trained with content scraped from databases where languages and norms of the dominant groups are statistically preponderant— minority groups are systematically prevented from contributing to filling those gaps. This epistemic asymmetry leads then to a *fracture* in the shared resources between mainstream knowledge on the one hand, reflecting the world-view of the dominant groups, and informing and shaping the online resources; and non-dominant knowledges and practices on the other, which, in addition to the mainstream knowledge, also include different kinds of awareness of minority norms and languages, of the gaps, the social realities and the injustices ignored by the dominant groups.

In this respect, then, the toxic deficiency of ML-based AI expresses not just a hermeneutical lacuna, but rather a form of *ignorance*. More exactly, a particular kind of ignorance that, prevalent among members of the dominant groups, is inherited by ML-based algorithms trained with content representing their (dominant) world-view. Moreover, this ignorance is not contingent, but rather a *systematic* feature of the shared online environment, produced and maintained as it is by the conformist attitude of AI design and training practices. And since it is an ignorance of the very oppressive systems that contribute to producing it, it is also not neutral, but plays an active role in upholding them, and in resisting its own erasure. Following Mills, then, I will refer to this *active* and *systematic* form of ignorance that contributes to sustain systems of oppression as a form of *white ignorance*.

In offering a characterisation of ML-based AIs as 'white ignorant', then, I propose to focus the attention on the following features of AI's toxic deficiency: *a*) its being part of

the very design of ML-based technologies, *b*) its active resistance to erasure, and *c*) its being undiscerning of non-dominant languages, norms and systems of meaning. If this is plausible, I want to show how, while, as an *hermeneutical lacuna*, the toxic deficiency of ML-based AI tends to impair minority groups as knowledge *seekers*, seen as a form of *white ignorance* it tends to obstruct them as knowledge *givers*.

To do so let's first go back to ASYLUM SEEKER. In this example, the AI is employed to evaluate the testimony of an asylum applicant against certain parameters and, by assessing their credibility, accept or reject their request. In the process of obtaining asylum, people who have been forced to leave their own country and have often suffered trauma and violence are put through the humiliating task of providing evidence of their conditions to the authority of the host country. Evidence of trauma, fear and violence, however, often cannot be other than testimonial —asylum seekers have to provide a story detailing the circumstances that have led them to flee their country. Because this story ought to be believed for the claim to be accepted, the success of the application depends on the accurate assessment of the applicant's credibility.

In recent years, a few countries (including Hungary, Latvia, Germany, Greece, Canada, the US and the UK) have been trialling the implementation of ML-based systems to carry out such assessments (Fair Trials 2021). Perhaps unsurprisingly, these practices have sparked huge controversy over the norms and criteria employed to generate the predictions. For instance, algorithms employed by the Home Office in the UK have been shown to take the applicant's nationality as a risk factor, and to rely on face recognition systems unable track cultural and racial differences, or the impact that traumatic experiences have on the way one reports them, both at the level of one's facial expressions and in the language and vocabulary employed (Fair Trials 2021, van den Hoven 2019, Eckenweiler 2019).

When asylum is denied on such grounds, it is precisely the machine's (white) ignorance of all these factors that causes it to fail to assign the right level of credibility to the applicant's

testimony. What I have in mind in this case, more exactly, is the machine's lack of resources apt to understand the system of meanings (such as the vocabulary and concepts as well as non-verbal cues and nuances of expression) of someone from a non-dominant background —like Negasi, for instance. In virtue of this lack, and owing to the prejudice ingrained in the machine, the categories applied by the algorithm to read and interpret Negasi's asylum application are inadequate to fairly assess the credibility of Negasi's testimony.

At the root of the injustice, then, a key role is played by a fundamental *communicative failure*. At bottom, that is, is the AI's failure to give a proper assessment of the applicant's credential that, in this case like many others, leads to the wrongful rejection of the applicants' request. Communicative failures of this sort, owing to the bias ingrained in a hearer's deficient conceptual resources, have been widely discussed in the literature on epistemic injustice. According to Kristie Dotson, for example, one can be a victim of a particular form of testimonial injustice (what she calls *testimonial quieting*) when a communicative failure is caused by a hearer's *pernicious* ignorance —that is, a kind of reliable ignorance that, in a particular context, tends to be harmful. More precisely, testimonial quieting involves cases where the pernicious ignorance of a hearer, in the form of negative stereotypes, or 'controlling images' (2011, 243), prevents them from perceiving the speaker as a knower, which causes them to fail to take up their communicative attempt. The communicative failure Dotson has in mind here is a form of illocutionary silencing, occurring when a hearer fails to take up a speaker's attempt to transmit a piece of information —for instance, when a woman's attempt to contribute to a conversation is taken to be a mere expression of her emotions[14]. In this case, the woman is said to be *silenced* because her utterance is not successful at being the kind of speech-act the woman intended it to be.

Thus understood, the epistemic violence of testimonial quieting bears intuitive similarity with the kind of injustice described in ASYLUM SEEKER. In both cases, the

[14] Case discussed in Tanesini (2016)

communicative exchange fails, and in both cases (systematic and wrongful) ignorance plays a key explanatory role. More exactly, in our case, it is the AI's ignorance, rooted in the machine's biassed functioning, that causes the algorithm to fail to assess the applicant' epistemic worth, ultimately leading to the communicative failure.

True, the testimonial exchange in ASYLUM SEEKER may not be considered strictly speaking *testimonial*, because it takes place between a human and a machine, and human-to-machine interactions do not obey the same norms as human-to-human —or so one may think. But it is at least not intuitively obvious why this should be a problem, at least with respect to the conversational norms relevant to this case. Indeed, it seems reasonable to expect that the conversational norm that is at stake here doesn't apply only to human communicative exchanges. After all, it is difficult to see how AIs could, say, give us the right predictions if they didn't recognise our requests as such —if they took, say, one's asylum request as a greeting. Even so, I do ultimately agree that it would be a stretch to subsume this case under the notion of testimonial quieting —at least in the way in which Dotson understands it. The reason is that the communicative collapse in this case does *not* involve a *failure of uptake*. Negasi's application has been *rejected*, which means that, at the very least, his speech act *is* acknowledged for what it is —i.e., an asylum request. If this is so, however, ASYLUM SEEKER does not describe a case of *illocutionary* silencing.

What's the issue in this case then? To a first approximation, I think that the problem can be understood as concerning the fact that the algorithm has prevented Negasi from obtaining the effect that, given their credentials, they were entitled to obtain with the communicative act they performed. If this is correct, the communicative failure at issue here does not concern *uptake* of the communicative act, but its *effect*. In other words, it is *perlocutionary* rather than illocutionary. The applicant has been perlocutionarily silenced: they have been unjustly prevented from obtaining something that they were entitled to obtain with their words (Spewak 2023).

When considered in their capacity as receivers of information, ML-based AIs are liable, owing to their active ignorance, to perlocutionarily silence members of minority groups' communicative attempts. The harm caused by having one's perlocutionary attempt frustrated is very common, and has recently been aggravated by the increased implementation of ML-based AI technologies. Studies by UC Berkeley, for instance, have found that Black people were consistently refused property loans due to the bias present in newly automated systems employed to process loan applications, which unjustly discriminated against applicants based on their ethnicity. Similarly, an investigation conducted by ProPublica in 2016 on the fairness of the criminal law system in Florida, has revealed that Black defendants' testimony were evaluated against an assessment of their likelihood to reoffend, which was in turn produced by ML algorithms that systematically discriminated against all non-White defendants.

These cases present patterns of injustice similar to the one in ASYLUM SEEKER, where a member of a minority group's attempt to obtain something through their communicative act is unjustly frustrated due to the systematic ignorance of ML-based AIs. Notice though how victims of AI-based perlocutionary silencing are clearly not *quieted*. Their communicative attempt doesn't go unacknowledged —instead, it is heard and taken up for what it is (a loan application, an asylum request, a non-guilty plea). The problem is rather that, in failing to obtain its goal, the attempt remains somewhat inert. Although it *is* heard, it is as though it wasn't. The Black woman who has applied for a loan, and whose request is being processed by the system, *has* been heard, and her communicative act has been taken up for what it really is —i.e., a request for a loan. Because it gets rejected, however, the request is unsuccessful, and she is unjustly prevented from obtaining what she had the credentials to obtain through that communicative act. Following this line of thought, then, we can say that the communicative attempts of victims of perlocutionary injustices, rather than being *quieted*, are unjustly shunned, or *spurned*. Expanding on Dotson's taxonomy, we can call *testimonial spurning* the kind of epistemic violence

occuring when active ignorance systematically silences the perlocutionary effect one is otherwise entitled to obtain with one's communicative act.

Looking at the toxic deficiency of ML-based AIs as a form of white ignorance then reveals a distinctive form of violence that, for the most part, targets members of minority groups in their capacity as knowledge *givers*. Following Dotson, I have proposed to think of this violence in terms of a communicative failure occurring when (white) ignorance causes one to fail to recognise the epistemic worth of their interlocutor. Departing from Dotson's analysis, and in an attempt at adding to it, I have suggested that, when it comes to theorising about ML-based forms of epistemic injustice more specifically, the communicative failure is better understood as concerning the perlocutionary effects of the speech act rather than its illocutionary force. Owing to this difference, I noted how the violence thus perpetrated concerns not the quieting as much as the spurning of one's testimonial attempt.

IV. CODA

In this paper I have done two things. First, I have looked at the design and training practices of ML-based AIs, and tried to show how this seems to present systematic flaws, and how these flaws appear to be, to some extent, the result of the implementation of a fundamentally mistaken principle regulating the machine's behaviour —what I called *epistemic conformism*.

Honing in on these results, I then tried to show how these design flaws impact AI users in their capacity as epistemic agents. In particular, looking at ML-based AIs in their function as gatekeepers of the knowledge stored in our shared online resources, I focussed my attention in particular on two basic epistemic aspects of the users' agency: their ability to seek and to pass on their knowledge. What I have found is that, with respect to both their knowledge seeking and knowledge giving abilities, ML-based AIs tend to set up barriers mostly affecting members of minority groups. The reason, I have argued, ought to be found precisely in the specific structure of the design flaws of

AI —particularly its toxicity and deficiency. More exactly, I have shown how the barriers erected against minorities' ability to give knowledge is connected to the white ignorance of AI, and how the barriers erected against minorities' ability to seek and obtain knowledge are connected to its hermeneutical lacunae.

If plausible, this seems to suggest a picture of ML-based AIs as systematically *ostracising* minority contributions. Considering the role that ML-based AI nowadays plays as gatekeepers of our shared online resources, and considering our increasing reliance on online content in our epistemic lives, the outright ostracism of minoritarian voices poses a serious threat to the integrity of our epistemic environments.

The growth of ML-based AIs, both in sophistication and extension of their application, is just at the beginning. The increase in implementation of ML-based technologies in everyday life is rapid and widespread. Since I started working on this article (in 2020, when my interest in machine learning was sparked by reading of the firing of Timnit Gebru from Google's ethics team[15]), the boom of AI has been exponential —in terms of the technologies that have been made available to the public (e.g., chatGPT or dall-e); in terms of the political and financial attention it has raised (e.g., more and more funding opportunities made available by governments all over the world to secure leadership in AI-related areas of research); and in terms of the critical attention it has raised (e.g., regarding worries about online assessments, or the fights over creative rights). Still, very little is being done to match this enthusiasm with sufficient critical examination. If anything, when we hear of Google's decimation of their ethics team, followed by Twitter and Microsoft's *en masse* suppression of theirs, the impression is rather that helpful criticism is being stifled.

Yet, I don't think that pessimism about the future of AI in our society is fully justified. We already have the conceptual tools and critical capacities to understand the threats posed by these new technologies and to improve them. Attention to the relationship between the ways in which we design and use AIs

[15] Hao (2020)

and issues of social justice is steadily increasing. New work (e.g., Huang, et al. 2022; Simion and Kelp 2023; Rafanelli 2022) is shedding light on possible solutions and virtuous models we can follow to develop better and more just AI. This paper should also be seen as an attempt in this direction. If I am correct, one optimistic stance is not justified: the one endorsed by those who take AI to be a neutral tool. According to this stance, the problem is not to be found in the functioning of AI itself, but in the way in which we make use of it. If I am right, we shouldn't find this stance fully satisfactory. For if, on the one hand, it is true that better training practices, as well as wider participation to online pools of data, may make for more virtuous AIs and alleviate some of our current worries, a solution to the problem requires more than that. And this is because the problem I have identified concerns AI's very design. The epistemic conformism of AIs is a design flaw which needs to be addressed directly. Failing to address this worry, I have argued, leads to distressing epistemic worries, like the epistemic marginalisation and ostracism of minoritarian perspectives.

REFERENCES

Adam, A. (1998) *Artificial Knowing: Gender and the Thinking Machine*, Routledge

Anderson, E. (2012) "Epistemic Justice as a Virtue of Social Institutions", *Social Epistemology*, Vol 26, No. 2, p. 163-173

Angwin, J. & Mattu, S. & Larson J. & Kirchner, L. (23rd of May 2016). "Machine Bias" *ProPublica,* online source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed Oct 2022)

Bagenstos, S. R. (2006). "The structural turn and the limits of antidiscrimination law" *California Law Review*, 94(1), 1–47.

Bender, E. M., Gebru, T., et al. (2021), "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?"

Proceedings of the 2021 ACM Conference on Fairness, Accountability, and TransparencyMarch. pp. 610–623, https://doi.org/10.1145/3442188.3445922

Browne, S. (2015), *Dark Matters*, Duke University Press.

Cave, S. & Dihal, K. (2020), "The Whiteness of AI", *Philosophy & Technology* 33, pp. 685–703. https://doi.org/10.1007/s13347-020-00415-6 (accessed Sept 29, 2021).

Du Bois, W. E. B. (1989). *The Souls of Black Folk*. Orig. ed. 1903. New York: Penguin.

Eckenweiler, L. (12th June 2019). "Seeking Asylum: Epistemic Injustice and Humanitarian Testimonies", in Justice in Global Health Emergencies & Humanitarian Crises (The University of Edinburgh): https://www.ghe.law.ed.ac.uk/seeking-asylum-epistemic-injustice-and-humanitarian-testimonies/ (accessed Oct 2022)

Fair Trials (2021). "Automating Injustice" online article sourced on: https://www.fairtrials.org/app/uploads/2021/11/Automating_Injustice.pdf

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

Friedman (2020), "The Epistemic and The Zetetic", The Philosophical Review 129 (4):501-536.

Gandy, O.H. (1998). *Communication and Race: A Structural Perspective*. Edward Arnold and Oxford University Press.

Hao, K (2018). "What is machine learning?" MIT Technology Review,

https://www.technologyreview.com/2018/11/17/103781/what-is-machine-learning-we-drew-you-another-flowchart/

Hao, K (2020). "We read the paper that forced Timnit Gebru out of Google. Here's what it says." MIT Technology Review, https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/

Hoffman, A. L. (2019) "Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse", *Information, Communication & Society*. Vol. 22, Issue 7: Data Justice.

Hornsby, J and Langton R. (1998), "Free Speech and Illocution", *Legal Theory*, 4 (1):21-37.

Huang, Linus Ta-Lun, Hsiang-Yun Chen, Ying-Tung Lin, Tsung-Ren Huang, and Tzu-Wei Hung (2022) "Ameliorating Algorithmic Bias, or Why Explainable AI Needs Feminist Philosophy." *Feminist Philosophy Quarterly*, 8 (3/4).

Jones, L. K. (2020). "Twitter wants you to know that you're still SOL if you get a death threat — unless you're President Donald Trump"
https://medium.com/@agua.carbonica/twitter-wants-you-to-knowthat-youre-still-sol-if-you-get-a-death-threat-unless-you-re-a5cce316b706 (access Oct 2022)

Kimmerer, R. W. (2013). *Braiding Sweetgrass*, Milkweed Editions

Krieger, L. H. (1995). "The content of our categories: A cognitive bias approach to discrimination and equal employment opportunity" *Stanford Law Review*, 47(6), 1161–1248.

Langton, R. (1993), "Speech Acts and Unspeakable Acts" *Philosophy & Public Affairs*, Vol. 22, No. 4, pp. 293-330.

Ledford, H. (2019), "Millions of black people affected by racial bias in health-care algorithms" *Nature*. Oct 29, 2019.

https://www.nature.com/articles/d41586-019-03228-6 (accessed Sept 29, 2021)

Longino, H. E. (1991) "Multiplying Subjects and the Diffusion of Power", *The Journal of Philosophy*, Vol. 88, No. 11.

Martín, A. (2021) 'What Is White Ignorance' *The Philosophical Quarterly* Vol. 71, No. 4

Medina, J. (2013), *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and the Social Imagination*. New York: Oxford University Press. Oxford Scholarship Online, 2013. doi: 10.1093/acprof:oso/9780199929023.001.0001.

Miller, J. (2020), "Is An Algorithm Less Racist Than A Loan Officer?" *The New York Times*. June 5, 2020. https://www.nytimes.com/2020/09/18/business/digital-mortgages.html (accessed Oct 29th, 2021).

Mills, C. W. (1997), *The Racial Contract*. Ithaca, NY: Cornell University Press.

Mills, C. W. (2017), *Black Rights/White Wrongs: The Critique of Racial Liberalism*. New York: Oxford University Press. Oxford Scholarship Online, 2017. doi: 10.1093/acprof:oso/9780190245412.001.0001.

Noble, S. (2018), *Algorithms of Oppression*, NYU Press.

Rafanelli, Lucia M. (2022), "Justice, injustice, and artificial intelligence: Lessons from political theory and philosophy" in *Big Data & Society* January–June: 1–5

Roser, M., Ritchie, H., Ospina-Ortiz, E. (2015) "Internet" *Our World in Data*, https://ourworldindata.org/internet (accessed Sept 29, 2021).

Simion, M., Kelp, C. (2023), "Trustworthy artificial intelligence" In *Asian Journal of Philosophy* 2, 8 (2023). https://doi.org/10.1007/s44204-023-00063-5

Snow, J. (2018, February 14). "We're in a diversity crisis": Cofounder of Black in AI on what's poisoning algorithms in our lives. MIT Technology Review. https://www.technologyreview.com/s/610192/were-in-a-diversity-crisis-black-in-ais-founder-on-whats-poisoning-the-algorithms-in-our/ (accessed Sept 29, 2021)

Spewak, D. (2023). "Perlocutionary Silencing: A Linguistic Harm That Prevents Discursive Influence" *Hypatia*, 1-19. doi:10.1017/hyp.2023.2

Spivak, G. C. (1999) *A critique of postcolonial reason: Toward a history of the vanishing present*. Cambridge, Mass.: Harvard University Press.

Thorstad, D. (2021) "Inquiry and The Epistemic", *Philosophical Studies* 178 (9):2913-2928.

Ullmann, S. (2021), "Google Translate is sexist. What it needs is a little gender-sensitivity training" *The Conversation*. Apr 05, 2021. https://theconversation.com/online-translators-are-sexist-heres-how-we-gave-them-a-little-gender-sensitivity-training-157846 (accessed Sept 29, 2021).

Willard-Kyle, C. (forthcoming) "The Knowledge Norm of Inquiry" *The Journal of Philosophy*.

van den Hoven, E. (2019). "Automated hermeneutical injustice", sourced online at: https://www.cohubicol.com/blog/automated-hermeneutical-injustice/#:~:text=Hermeneutical%20injustice%20according%20to%20Fricker,tropes%20that%20operate%20within%20society.