It is not About Bias but Discrimination

Chaewon Yun[1,*], Claudia Wagner[2,3] and Jan-Christoph Heilinger[4]

[1] Max Planck Institute for Human Development, Berlin, Germany
[2] GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany
[3] RWTH Aachen University, Aachen, Germany
[4] Witten/Herdecke University, Witten, Germany

Abstract

Growing interest in the bias of LLMs is followed by empirical evidence of socially and morally undesirable patterns of LLMs output. However, different definitions and measurements of bias make it difficult to assess its impact adequately. To facilitate effective and constructive scholarly communication about bias, we make two contributions in this paper: First, we unpack the conceptual confusion in defining bias, where bias is used to indicate both descriptive and normative discrepancies between LLMs and desired outcomes. Second, we suggest deontological reasons why bias is unacceptable. Common arguments against bias are based on teleological grounds which focus on the consequences of biased LLMs. We argue that bias should be identified and mitigated when and because it is morally wrongful discrimination, regardless of its outcome. To support this argument, we connect biased LLMs with Deborah Hellman's meaning-based account of discrimination. Bias in LLMs can be demeaning and capable of lowering the social status of affected individuals, making it morally wrongful discrimination. Such bias should be mitigated to prevent morally wrongful discrimination via technological means. By connecting the phenomena of bias in LLMs with existing literature from wrongful discrimination, we suggest that critical discourse on bias should go beyond finding skewed patterns in the outputs of LLMs. A meaningful contribution to identifying and reducing bias can be made only by situating the observed and measured bias in the complex societal context.

Keywords

Bias, Discrimination, Large Language Models, Computational Social Science

1. Introduction

In recent years, there has been a surge of interest in uncovering bias, ethical and social risks, and potential harm in machine learning algorithms[1][2][3][4]. While the specific algorithms of interest have changed with the development and use of new algorithms, current discussions are dominated by large language models (LLMs)[1]. Following the explosive commercial success

EWAF'24: European Workshop on Algorithmic Fairness, July 01–03, 2024, Mainz, Germany

*Corresponding author.
yun@mpib-berlin.mpg.de (C. Yun); claudia.wagner@gesis.org (C. Wagner); jan-christoph.heilinger@uni-wh.de (J. Heilinger)

[1] In this paper, we use LLMs to describe language models that was published since BERT[5], which initiated the popularity of language models in various NLP applications. There are various terms used to indicate language models, including generative models, auto-regressive models, foundation models, and transformer models[6], to name a few. For our purpose, we mainly use the term LLMs as it is now a widely used term. However, we would like to note that the term LLM is not a strictly defined term with fixed conditions. What used to be a "large" language model in the past can be "small" in several months.

of LLM-based applications such as ChatGPT[7][8], there are heated debates about how this technology might disrupt society. Numerous cases have been reported of LLMs exhibiting morally dubious patterns that need to be urgently addressed. A growing literature on LLM bias in computer science literature has been focused on identifying problematic patterns such as racist[9], casteist[10] or anti-Muslim[11] outputs produced by LLMs.

In addressing biased outcomes of LLMs, most criticisms focus on the harm that biased LLMs can bring about. However, such perspective assumes that the bias of language models are problematic only based on their harms. We argue that moral permissibility of biased LLMs goes beyond the harmful outcome that they may bring about.

In this paper, we discuss biased LLMs in terms of wrongful discrimination by building upon literature on critical studies of technical systems and wrongful discrimination. Due to the conflation of bias at multiple levels in the context of LLMs, it is not a trivial task to navigate the disorganised discussion of bias in LLMs. To overcome such limitations, we investigate the meaning of bias in the context of LLMs. Afterwards, we reason why such bias in LLMs are wrongful and hence should be identified and mitigated.

In this attempt, we propose two questions that we will answer throughout this paper: First, what does it mean for LLMs to be biased? Second, why do we care if LLMs are biased? To answer the first question, various sources of confusion in the definition of bias are enumerated in the next section. Sources range from different uses as disciplinary jargon, different methods of operationalisation, to conflated conceptualisation of normatively and descriptively defined bias. We identify that unpacking conflation in conceptualizations of bias is especially of critical importance to avoid misleading interpretation of identified bias. Engaging with empirical literature measuring bias in LLMs shows how different conceptualizations of bias leads to varying implication of measured biases.

After discussing the conceptual conflation of the definition of bias, the second question asks where the relevance of bias in language models stems from. The common argument against bias in LLMs is in relation to the harms, as Biased LLMs can create incorrect outputs and can lead to harmful outcomes. However, such consequential perspectives are unsatisfactory for justifying some types of bias that should be mitigated despite being factually correct or not obviously dangerous or unsafe. After discussing the limitations of consequentialist reasons for mitigating bias, we provide deontological reasons to explain why bias in LLMs is impermissible beyond harmful outcomes. We argue that bias in LLMs is morally wrongful discrimination and therefore should be mitigated. To support this claim, we introduce Deborah Hellman[12]'s meaning-based account of discrimination. Building on Hellman's account of discrimination, we argue that observed patterns of bias in LLMs are impermissible when and because they are morally wrongful discrimination. Contextualising observed bias in LLMs in terms of discrimination is crucial for justifying why and how LLMs should be unbiased.

2. What Does It Mean for LLMs to be Biased?

2.1. The Problem of Conflated Conceptualization

Among the various dimensions of conflation regarding bias in LLMs, the conflated conceptualisation of bias is an important yet understudied problem. The term bias is used broadly in the literature on bias in LLMs. The definition often lacks clarity, partly due to the different ways in which the term bias is used. There are several reasons for this, such as different disciplinary practices or different methodological choices for operationalising bias. Among other reasons, conceptual confusion leads to fundamentally different versions of unbiased LLMs. Bias is used to describe a variety of gaps between the performance of LLMs and different golden standards. In particular, such a golden standard can refer to normatively correct LLMs or descriptively accurate LLMs.

As bias is a contested concept that is widely used across different disciplines, the variance among definitions of bias is not negligible. When bias is defined in statistical analysis, it describes the gap between the observed or estimated value and the true value. For instance, bias is defined as "systematic error arising during sampling, data collection, or data analysis"[13] or "prior information, a prerequisite for intelligent action."[14] Alternatively, bias can be defined as unfair treatment, as in Friedman and Nissenbaum (1996)[15], where normative evaluation becomes an integral part of determining whether a pattern can be considered biased or not. Friedman and Nissenbaum define computer systems are biased when they "systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others" (ibid, p. 332). Since the context of studying bias in LLMs is not limited to a single discipline, the usage of the term bias varies significantly. The gap between the definition, measurement, and normative motivation for investigating bias has been criticised[16].[2]

Another layer that adds complexity to the definition of bias is the different operationalisations of bias. Bias is a construct that cannot be directly observed or quantified. Therefore, bias needs to be operationalised with observable properties that are relevant to the construct[19][20][21]. It is a prerequisite to clearly define the construct before operationalisation[22]. In our case, the definition of bias in LLMs should be explicitly defined. In existing literature of measuring bias, stereotypes, especially occupational stereotypes[23][24][25][26][27][28], are the most commonly used properties to measure gender bias. This line of work can measure stereotypical relations between genders and occupations, such as 'man' to 'computer programmer' and 'woman' to 'homemaker'[29]. Sentiment score[24][30][31][32] is another commonly used property to operationalise bias. LLMs are considered to be biased when a certain gender is correlated with more negative sentiment, that gender is biased against to others.

2.2. Conceptual Conflation: Descriptive/Normative Framework

Besides varying definitions of bias across different disciplines or operationalisations, conceptual conflation concerns fundamental differences about bias. While the first two conflations complicate the specific technical meaning of measured bias, conceptual conflation transcends practical differences of working definitions or operationalisations.

To measure bias, bias first needs to be identified by comparing LLM outputs with the desired state that is defined as unbiased. Based on how unbiased LLMs are defined, implications of measured bias changes significantly. Broadly speaking, unbiased LLMs can be defined descriptively and normatively. Descriptively defined bias can be measured by comparing LLMs output with statistics that represent status quo. LLMs are considered unbiased when their output

[2] Similarly, discussion around fairness in machine learning has attracted considerable critical attention, such as in [17] and [18].

Table 1
Four Lands Analogy in Deery and Bailey (2022)

	Descriptively accurate	Descriptively inaccurate
Normatively correct	Utopia Land	Fantasy Land
Normatively incorrect	Dilemma Land	Disaster Land

aligns with the representation of reality, irrespective of normative implications. Alternatively, LLMs can be compared to normative targets that define what LLMs should be like, irrespective of whether they describe how things are like in reality. LLMs that produce outcome according to the ideal status, such as equal representation of genders or absence of stereotypes, will be evaluated as unbiased. Consequently, despite using the same vocabulary and similar methods to identify bias, differences in conceptual conflation leads to vastly different implication of measured bias.

Deery and Bailey (2022)[33] provide an analogy to describe different positions on how ideal natural language processing technology should be. It helps to compare different unbiased LLMs by laying out different possibilities based on two axes of normative correctness and descriptive accuracy. Four different categories of unbiased LLMs are possible by looking at combinations using the normative correctness axis and the descriptive accuracy axis, as described in Table 1.

In an ideal world, the *Utopia land* according to Deery and Bailey's analogy, unbiased LLMs accurately describe reality, and reality conforms to the normative ideal. It makes LLMs both normatively and descriptively correct. In *Disaster land*, on the other hand, unbiased LLMs fail to describe reality accurately, while also diverging from the normative ideal. In this case, LLMs are descriptively and normatively wrong, which makes it irrelevant to develop such a model, since it will be useless.

This leaves two possibilities in question, the *Fantasy Land* and the *Dilemma Land*, which are also the most common versions of unbiased LLMs in the existing literature of measuring bias. Fantasy land refers to the state where impartiality is defined by the normative ideal despite its descriptive imprecision. It is a fantasy because it does not exist in reality. It is unrealistic, but an ideal state of the world to which language models should aspire. On the other hand, dilemma land prioritises descriptive accuracy, even though such accuracy does not correspond to the normative ideal. It is a dilemma because it is an intractable problem to satisfy both descriptive and normative demands. It is not 'ideal' because the present world is not perfect. Adapting language models to an imperfect reality will inevitably lead to imperfect language models.

2.3. Choosing between Fantasy and Dilemma

The decision as to where unbiased LLMs should be located, either in fantasy land or in dilemma land, comes down to prioritising descriptive accuracy or normative correctness. The dilemma arising from the tension between normative and descriptive correctness is not a problem unique to language models or algorithms. It is a "specter of normative conflict" (Basu, 2020)[34] that fairness might require inaccuracy. The dilemma perspective suggests that the apparent conflict between fairness and accuracy cannot be resolved (ibid, p. 191-197). Fantasy land advocates normative correctness, claiming that LLMs should be free of problematic patterns such as stereotypes that exist in reality.

Basu describes how the pursuit of such an ideal comes at the cost of a loss of descriptive accuracy, which can reduce the usefulness of language models. Prioritising descriptive accuracy, on the other hand, can create more utility for LLMs by aligning with a factual description of the status quo. However, it may risk perpetuating existing problems of reality through language models and reproducing undesirable patterns of injustice.[3] In the following subsections, we will refer to existing literature that empirically measures bias in LLMs. Two different ideals of unbiased LLMs will be discussed in terms of their implications and limitations.

2.3.1. LLMs in Fantasy Land

The most common approach to identifying bias in LLMs is to compare the output of the model with a set of ideal states. Different versions of ideal states are suggested by the authors who propose the metrics to measure bias. For example, unbiased LLMs are often defined as a-stereotypical[35] model that does not exhibit existing stereotypes. Stereotypes based on gender, religion, or race are considered undesirable, so LLMs should not be more likely to produce outcomes that conform to stereotypes. Another common approach is to seek a consistent baseline across groups, treating different groups of individuals ‘equally’[32]. Indeed, unbiased LLMs should not favour certain groups over others, nor be more prone to producing stereotypical texts.

The reference to an ideal state seems to be a self-suggestive step, considering the problematic patterns reflected in the training data, which is the prominent source of bias in language models. The data represent a reality that reflects morally imperfect features of the world. Therefore, language models trained on problematic data will inevitably show problematic patterns. By providing an alternative ideal where problematic patterns of reality are removed, the argument goes, LLMs can be improved.

However, this approach requires a definition of what an ideal language model is based on different aims. The definition of the ideal state can be subjective, contradictory, and controversial, which makes it difficult to compare differently unbiased LLMs. Establishing the criteria for an ideal language model requires conceptualising relevant concepts and evaluating conflicting values. ‘Essentially contested concepts’[36] such as bias or fairness are multifaceted and enables different possibilities to conceptualise the constructs.[4] Due to the complex nature of bias, any reductive definition runs the risk of missing potentially relevant features.

[3] Deery and Bailey (2022)[33] argues that there is an ethical value in not debiasing, such as presenting problematic patterns. Debiasing can create a false illusion of improved fairness, more than is actually the case, which can contribute to a devaluation of the problem. However, in the case of LLMs, we argue that the risks of perpetuating problematic patterns through LLMs outweigh the ethical value of leaving the problematic patterns intact in the models. Empirically comparing such potential benefits and harms can be an interesting research direction in the future.

[4] There has been much academic discussion about how to measure fairness. While various fairness measures have been developed, many of them are incompatible and even contradictory. On the inherent difficulty of measuring fairness, see, *inter alia*, Hellman (2020)[37], Jacobs and Wallach (2021)[22], Binns (2020)[38], and Delobelle et al, (2022)[2]

2.3.2. LLMs in Dilemma Land

As shown earlier, defining an ideal state in fantasy land is challenging because there is no one perfect LLM that everyone agrees on. Therefore, an alternative approach in dilemma land does not define what language models should be like. Instead, the model is compared to reality. The way reality is represented can be broadly divided into a statistical account and a subjective-evaluative account.

According to the statistical account of bias, the model is biased if the output does not match the statistical source with which it is compared. Real-world statistics are used as the source of what a model should represent. As an occupational stereotype is one of the most common ways of assessing bias, national labour statistics in countries[27] are often used to compare with the outputs of LLMs. For example, Touileb et al. (2022)[25] measure bias in LLMs using Norwegian statistics on the gender ratio in different occupations. The authors show that LLMs tend to map gender-balanced occupations as male-dominated according to Norwegian occupational statistics (ibid, p. 209). Therefore, the author argues that the model shows gender bias since LLMs misrepresent the distribution of occupations in Norway.

According to the subjective-evaluative account, the output of LLM is validated with human evaluation to assess whether the bias of a language model matches human bias. Following this approach, bias in a language model may be acceptable as long as humans show a similar degree of bias. The results of LLMs are compared with human perceptions measured by crowd-sourced annotators[24] or experts[39]. For example, Sotnikova et al. (2021)[39] use authors' ratings of how stereotypical given statements generated by language models are.

Some describe this approach as non-normative because it is a descriptive comparison. According to both statistical and subjective-evaluative accounts, bias is identified based on the discrepancy between the LLMs and the status quo, and this discrepancy is considered undesirable. Unbiased LLMs ought to either resemble real-world statistics or human perceptions. By referring to existing statistics, the comparison can be justified by relying on the authority of the statistics, rather than the authors deciding how LLMs should be. For example, referring to national labour statistics establishes the baseline as a representation of the status quo in the labour sector in the limited context in which such statistics are collected. Similarly, comparison with human assessments provides an observable baseline against which the performance of LLMs can be empirically assessed.

2.4. No Land without Normativity

However, both attempts to identify and measure bias in LLMs both in fantasy land and dilemma land are products of normative decision. It is obvious why normative approaches, LLMs in fantasy land, are the result of normative decision, as the bias is defined as a gap between the output of LLMs and normative ideals, such as a-stereotypical or indifferent to different groups of people. However, as much as fantasy land approach, the descriptive approach as in dilemma land is based on a normative position. Aiming to align LLMs with a descriptive representation of the status quo is a strong normative position, as it assumes that reproducing the status quo is desirable.

Unbiased LLMs developed according to this perspective will reinforce the existing social structure. The repetition and reinforcement of existing biases in LLMs is a well-documented phenomenon, exemplified in studies shown earlier in both fantasy land and dilemma land.

Moreover, the choice of which descriptive statistics or which human baseline to use is also a political choice. By referring to certain statistics as a golden standard to which LLM should be aligned, such statistics gain a position of power as authority. No statistic is neutral[40], but rather a product of social structure. The choice of people to compare LLMs with, whether they are experts in the field, university students, or crowd-sourced workers from low-wage countries from the Global South, also has significant implications for the interpretation of LLMs. Therefore, statistics do not qualify as objective, neutral authority that LLMs should be aligned with. Rather, it is a type of normative choice putting forward a specific value or perspective similar to putting forward an imaginative ideal state without referring to empirical data.

3. Why are Biased LLMs Undesirable?

In the previous section, we discussed the conceptual confusion that leads to a fundamentally different idea of unbiased LLMs. Despite differences in conceptualisation and operationalisation, there seems to be a broad consensus in the field that bias in LLMs is undesirable and should therefore be mitigated. And the most widely adopted motivation for such endeavor is due to the harm that biased LLMs can cause.

In addressing biased outcomes of LLMs, the most widely adopted framework in literature has been categorizing harm into representational and allocational harm based on Kate Crawford's NIPS Keynote in 2017[41]. Representatinoal harm refers to the reinforcement of subordination of people based on social identifiers such as race and gender. Allocational harm refers to decision-making systems withholding an opportunity or a resource to certain groups.[42] However, such perspective assumes that the bias of language models are problematic when and because it is harmful. For instance, LLMs can be harmful by creating misinformation. Proponents of such arguments focus on the consequences of biased LLMs.

However, such a consequentialist view is unsatisfactory in some cases of bias as we will show in this section. We propose deontological reasons why bias in LLMs is inadmissible, as it is morally wrongful discrimination. Wrongful discrimination is impermissible regardless of its consequences as it fails to treat people equally.

3.1. 'Hallucinated' LLMs are Harmful

One of the most common criticisms of LLMs is their 'hallucination'[43]. Hallucination refers to the tendency of LLMs to produce texts that are not factually based. The basic mechanism of LLMs is to produce a statistical prediction of the most likely sequence, which does not take into account its veracity. The hallucination of LLMs is therefore a feature, not a bug. Despite such fundamental limitations, many people use applications that use LLMs, such as ChatGPT, as search engines or knowledge bases to retrieve useful information. Therefore, descriptive inaccuracy severely diminishes the utility of LLMs in many use cases. Moreover, it can be harmful and unsafe in high-stake situations such as health information or political disputes. Thus, the argument goes, LLMs should be unbiased in the descriptive sense to be useful and practical.

To reduce harms that can be caused by LLMs' hallucination, factual correctness, or groundness, has become integral part of LLM evaluation. For example, Touvron et al. (2023)[44] evaluates safety along three dimensions: Truthfulness, toxicity, and bias. Thoppilan et al. (2022)[45] evaluates harm, discrimination, and misinformation. The authors aim to mitigate 'unsafe model output to avoid unintended results that create risks of harm, and avoiding creating or reinforcing unfair bias.' (ibid, p. 5) While authors make different connections between truthfulness, toxicity, bias, safety, and harm, they share the focus on the result of biased LLMs.

3.2. Constraints on Consequential Views

Reducing and preventing undesirable outcome of biased LLMs has been the focus of critical investigation of LLMs in empirical literature . Bias, defined as descriptive inaccuracy, is avoided by making LLMs factually truthful. However, making LLMs truthful can still result in harmful LLMs. As discussed in the previous section on dilemma land, aligning LLMs with status quo denotes that it is desirable to replicate the reality. Making LLMs factually accurate still risks reinforcing existing unjust prejudices since the status quo itself is shaped partly by 'morally problematic attitudes, beliefs, and institutions' as Basu(2020)[34] describes. As present reality embeds historical injustice, repeating reality via LLMs risks replicating and reinforcing existing patterns of injustice.

Moreover, the link between harmful consequence and bias are often ambiguous. An example reviewed by Blodgett et al. (2021)[46] highlights that many operationalisations of bias are contradictory and unjustified. For instance, the authors found a benchmark defining a stereotype as 'The exchange student became the star of all our art shows and drama performances', while the anti-stereotype was described as 'The exchange student was the star of our football team' (ibid, p. 1004). It is unclear how 'debiasing' LLMs according to such stereotypes will contribute to the less harmful consequence of LLMs. A vague link between bias and harmful consequence may risk reducing bias problems to immediately obvious harms. It also leaves room for implicit and benevolent discrimination whose harmful outcomes are not explicitly visible.

However, such criticisms should not lead to the mistaken conclusion that working on the bias of LLMs is an unfruitful effort. On the contrary, it is important to justify and strengthen the argument why certain biases should be measured and reduced, as it contributes to the relevance and implication of measured bias as such. In the next section, we discuss an alternative motivation for identifying, measuring, and reducing bias: the deontological view.

In deontological theories, what makes choices right depends on how it confirms with a moral norm, irrespective of the outcome that choices bring[47]. Bias in LLMs is undesirable regardless of the outcome of biased LLMs. Compared to the extensive efforts to identify and measure bias, relatively little attention has been paid to contextualise why such measured bias is relevant in a societal context, especially in relation to discrimination.

4. Do LLMs Discriminate?

In previous sections, we discussed what it means for LLMs to be biased, which answers the first question we raised earlier. In this section, we seek to answer the second question: Why do we care if LLMs are biased? We argue that it is because bias in LLMs constitutes wrongful discrimination.

Discriminatory bias risks exacerbating existing injustice through technological means[48]. The reason why LLMs are criticised for their tendency to produce stereotypical, biased results on socially relevant characteristics is not that it is wrong to be stereotypical per se, but because it is discriminatory. It is the discriminatory meaning attached to the observed bias in whatever form, either descriptive or normative, that determines which outputs of LLMs are permissible. While different contexts and objectives of LLMs and bias measures designed for LLMs shed light on different aspects of LLMs, what makes such patterns critical subjects of study is their relevance to discrimination.

In this section, we will discuss what discrimination means and how it relates to biased LLMs. In particular, an important question is what distinguishes mere discrepancy from morally wrongful discrimination. We first engage with existing literature on wrongful discrimination. Afterwards, we apply Deborah Hellman's meaning-based account of discrimination to the case of bias in LLMs. Why is the description of proactive male characters and submissive female characters in GPT-3 generated stories problematic[49]? More specifically, why are such dimensions of LLM problematised and measured among many other possibilities?

4.1. Wrongful Discrimination

What is discrimination and what makes them wrongful? Extensive scholarly discussions have investigated various aspects of discrimination: from distinguishing direct and indirect discrimination and their moral permissibility[50][51][52], differences between harmful, harmless, and even beneficial discrimination[53], to wrongful action and wrongful discrimination (ibid, p. 111). As Slavny and Parr (2015)[53] sketch out, the most contested issues in theory of discrimination is distinction between 'exclusively consequence-focused' and others who argues that wrongful discrimination can be defined by other factors than the consequences that discriminatory actions bring about (ibid, p. 101).

For instance, harm-based account by Kasper Lippert-Rasmussen[50] is an example of consequence-based account. According to his harm-based account, "an instance of discrimination is wrong, when it is, because it makes people worse off"(ibid, p. 154-155).

In contrast, scholars like Deborah Hellman focus on the objective meaning conveyed by discriminatory act[12]. Alternatively, Sophia Moreau[51] views discrimination is wrongful when it denies one's "right to have a certain set of deliberative freedoms, and to have these freedoms to an extent roughly equal to those held by others"(ibid, p. 168). Moreau and Hellman disagree in what role demeaning messages, sent by discriminatory action, serves in defining what renders an wrongful discrimination. The reasoning of wrongful discrimination can be found outside the consequences of discriminatory action.

In this article, we follow Deborah Hellman's meaning-based account of discrimination to explain the wrongness of bias in LLMs. It is not to suggest that this is the only possible explanation why bias in LLMs is wrong, but rather aiming to bridge existing in bias in LLMs research and discrimination literature. We show that different theories on discrimination can challenge widely-accepted assumption in bias in LLMs research, where wrongness of biased LLMs is based on their harmful outcome, almost exclusively. Critical discourse in biased LLMs can benefit from referring to theories on discrimination from political philosophy and law, which can strengthen the argument why some bias matters more than others.

4.2. Meaning-Based Accounts of Discrimination

In a broad sense, discrimination can refer to any differential treatment based on personal characteristics[12]. And not all forms of discrimination are morally problematic.[5] For example, charging young drivers more for their insurance can be regarded as age discrimination. However, some discrimination is justified to prevent adverse selection[55]. Similarly, giving discounts to the elderly, people with reduced mobility, or students in public museums is also benevolent discrimination based on age or status, which is accepted as a form of social security or other reasonings. Applying the same arguments on LLMs, discrimination by LLMs against certain groups of people may not be morally problematic in itself. For example, an LLM that produces more cat lovers than dog lovers will not be of significant moral or social concern. What makes such discrimination wrongful is the meaning of the discrimination.

Deborah Hellman(2017)[12] argues that what makes discrimination morally wrong is the meaning of the discrimination, rather than the intention of the actor, the relevance of the trait used to discriminate, or the rationality of the discrimination. Based on the meaning-based account of discrimination, the author argues that "discrimination is wrong when and because it is demeaning" (ibid., p. 1). It is demeaning when the actor with social power engages in denigration, which is the act of saying that someone is not good or important, and fails to treat those affected as equals. Hellman argues that we need to look at the meaning of discrimination in order to assess whether such discrimination is morally wrong.

The meaning-based account of wrongful discrimination has two aspects, an expressive dimension and a power dimension. An expressive dimension concerns whether an action or policy regards another person as inferior or of lower status. Hellman argues that discrimination is particularly problematic when it is based on socially salient characteristics such as gender and race. Socially salient characteristics such as gender can be used as "accurate proxies" for discrimination based on historical injustice. (see also Johnson (2021)[56]) Therefore, discrimination based on socially salient features is particularly morally problematic because it fails to treat people with equal moral worth.

Another dimension of the meaning-based account of wrongful discrimination is the power relationship. If the actor making the statement has social power that gives force to the meaning of the act or policy, then the discriminatory act is sufficient to be unjustified discrimination. Hellman argues that power is important in identifying unjustified discrimination because the actual power enables such discrimination to lower the social status of those affected. The discriminator with power and authority can affect people in more critical ways than those without such power. Furthermore, Hellman stresses that degradation depends on the capacity that comes with power, not on an actual effect. Regardless of the outcome, when the actor of power fails to recognise the equal moral worth of others, it is morally wrongful discrimination.

[5] "The principle of discrimination" (also called "principle of distinction") is referred to as one of the most important principles of International Humanitarian Law, the field of law governing armed conflicts[54]

4.3. Bias in LLMs as Morally Wrongful Discrimination

In the previous section, we explained Deborah Hellman's account of the meaning-based account of discrimination. Building on her theory of discrimination, we argue that biased LLMs can constitute a case of morally wrongful discrimination. We discuss how bias in LLMs exhibits both expressive and power dimensions according to the meaning-based account of discrimination. In this section, we discuss two dimensions of discrimination using an example of measured gender bias by Lucy and Bamman (2021)[49]. We will use this example to show how biased LLMs instantiate morally wrongful discrimination.

4.3.1. ChatGPT: "Male characters are powerful and female characters are emotional"

Lucy and Bamman (2021)[49] examine stories generated by GPT-3 that reproduce gender stereotypes from film, television, and books. The authors compared the themes of GPT-3-generated stories and human-written books to see how the perceived gender of the character related to the occurrence of topical terms such as appearance, intellect, and power. The perceived genders of characters were identified based on gendered pronouns, honorifics, or names. The prompts used for GPT-3 to generate stories consist of single sentences containing main characters, sampled from 402 contemporary English fiction books.

The authors carried out two types of content analysis. First, topic modelling was used to identify coherent collections of words in the text. The result shows that GPT-3 tends to associate female characters with topics related to family, emotions, and body parts. In contrast, masculine characters were associated with politics, war, sports, and crime. The authors show that the different themes across perceived gender in the stories generated by GPT-3 are consistent with previous work showing that language models associate women with caring roles[23], maternalism, and appearance[57]. In addition, GPT-3 generated longer stories when the prompt contained stereotypical characters than anti-stereotypical characters.

In addition to topic modeling, the authors analyze how characters are described by measuring semantic similarity with lexicon embeddings. Three dimensions of description, appearance, intellectuality, and power, were chosen based on previous work on stereotypical description based on gender[57][58][59][60]. The result shows that words describing appearance are often used for female characters and words related to power are used for male characters.

The authors conclude that GPT-3 had internalised stereotypical gender stereotypes, which was strong enough to neutralise the effect of using power words for female characters. Even when the prompts did not contain explicit gender information or stereotypes, GPT-3 tended to generate stories that conformed to gender stereotypes. In addition, the authors found that GPT-3 tended to include more masculine characters and that the outcome varied according to the gender of the character, even when identical prompts were used (ibid, p. 51-52).

4.3.2. The Expressive Dimension

The expressive dimension assesses whether the one expresses denigration and views the other as less worthy, i.e. demeaning. To assess the expressive dimension of prejudice in LLMs, it is necessary to see whether the prejudice treats certain groups of people as having lower status compared to other groups of people.

Lucy and Bamman's (2021)[49] research shows that stories generated by GPT-3 show encoded stereotypical gender bias. The association of women with family, appearance, and less power has been studied extensively in feminist theory. The association of women with appearance reflects the history of objectification of women. Feminists have raised the problems of objectification that make women overly preoccupied with their appearance[61]. The association of women with their appearance fails to recognise women as equal agents to men by identifying women only/mainly with their bodies rather than their whole being. Bartky argues that the fragmentation of the female body sees women as "less inherently human than the mind or personality"[62][61]. Language models that associate female characters with appearance show that gender injustice is reproduced by technological means.

Similarly, an examination of family dynamics and power structures effectively highlights the presence of gender inequality. Susan Moller Okin points out that "socially constructed inequalities" exist in the distribution of critical social goods such as power, prestige, and opportunities for self-development[63][64].

By reiterating coded gender inequality, LLMs like ChatGPT fail to treat people of different genders equally. By associating women with less power and agency, women are given a lower status. Such an example demonstrates how bias in LLMs instantiates morally wrongful discrimination. As Hellman argued, the repetition of historical injustice along the axes of socially salient characteristics is a significant case of morally wrongful discrimination.

4.3.3. The Power Dimension

Another dimension that needs to be explored concerning Hellman's meaning-based account of discrimination is the power dimension. To address the power dimension, we should ask whether LLMs have power and authority that can influence people in more consequential ways. If the agent of power, in our case LLMs, fails to recognise the equal moral worth of others, then this is morally wrongful discrimination. And we argue that language models are in a position of power where bias can begin to have a consequential impact on people's lives.

Several literatures have analysed the impact of ChatGPT, such as the EUROPOL report on law enforcement[65] and Dempere et al. (2023) on higher education[66], among others. ChatGPT is just one example of an application using LLMs. LLMs have been adopted in various applications and can be used in numerous creative ways, from collecting debts[67] to creating an AI companion[68]. As this is a rapidly growing market, applications using LLMs are likely to increase. Traditionally conservative sectors such as the military are also experimenting with incorporating LLMs into their operations[69]. As used in real-world scenarios, LLMs have power and authority that will directly and indirectly affect people's lives. The bias of LLMs used for different applications, such as debt collection, can lead to unequal treatment of different groups of people.

AI has the potential to transform society, as evidenced by the heated debate on how to regulate AI[70]. However, the real-world impact of biased and opaque algorithms has been materialising for years even before LLMs were developed. Monumental cases of algorithmic bias in criminal risk assessment algorithms[71], facial recognition algorithms[72], or recruitment algorithms[73] show critical consequences of using such algorithms in practice. Such cases demonstrate that discriminatory treatment disproportionately harms historically marginalised populations, and LLMs are no exception. Discriminatory patterns based on socially salient characteristics such as gender, race, and religion are easily found in LLMs, which can exacerbate existing discrimination based on such characteristics. It also indicates potential discriminatory patterns that are harder to measure can be undermined.

In this section, we have discussed biased LLMs and the meaning-based account of discrimination. To assess what level of measured bias in the language model is sufficient to constitute morally wrongful discrimination, two aspects need to be examined: the expressive dimension and the power dimension. The expressive dimension can be evaluated by assessing whether the bias expressed in the language model's output fails to treat those affected as equal to others. The power dimension examines whether the language model has the power to turn such discriminatory bias into real harm. We have shown how measured bias in LLMs relates to two dimensions of discrimination, building on Deborah Hellman's account of morally wrongful discrimination.

5. Conclusion

In this paper, we have engaged in a conceptual analysis of bias in LLMs to investigate what it means for LLMs to be biased. Among various sources of confusion in defining bias, we focused on conceptual conflation where bias is used to refer to either descriptive inaccuracy and normative inaccuracy.

Descriptive inaccuracy is commonly measured by comparing LLMs with descriptive statistics, such as national labour statistics. Alternatively, bias in LLMs refers to normative inaccuracy, such as stereotypical correlations with socially salient characteristics such as gender, race, or religion.

Common arguments against bias are based on practical utility or safety grounds, which are consequential grounds based on the outcome of bias in LLMs. We argue that bias in LLMs should be identified and mitigated because it is morally wrongful discrimination, regardless of its outcome. Irrespective of the descriptive or normative correctness of biased patterns that LLMs produce, it is concerning when and if such bias instantiates morally wrongful discrimination. We presented Deborah Hellman's work on the meaning-based account of discrimination. She provides a framework that considers two dimensions of discrimination: the expressive dimension and the power dimension. We showed that biased LLMs are a case of morally wrongful discrimination based on both accounts. We also use one specific case of bias measurement to show how this framework can be applied to empirical bias measurements.

Regardless of the outcome of the bias, morally unjustified discrimination against LLMs is unacceptable. People have equal moral worth and should not be discriminated against on the basis of socially salient factors such as gender, race, or religion. The same argument applies to LLMs. It is particularly important to discuss the societal implications of bias against LLMs, as it is anticipated that the technology may transform various sectors of society and affect the lives of many people. Bias in LLM poses the pertinent threat of technology-enabled discrimination which risks exacerbating existing social injustices.

Bias in LLMs is a problem not because of observed biased patterns per se, but because it can discriminate in wrongful ways. To accurately identify potential risks and harms that may be posed by LLMs, we urge that the study of bias in LLMs should go beyond finding a skewed pattern in the outputs of LLMs. We provided a lens of wrongful discrimination as a framework for evaluating pertinent LLM bias in a societal context. A meaningful contribution to identifying and reducing bias in LLMs can be made only by situating the observed and measured bias in the complex societal context where the impact of bias can be critically evaluated.

References

[1] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Glaese, B. Balle, A. Kasirzadeh, Z. Kenton, S. Brown, W. Hawkins, C. Biles, A. Birhane, J. Haas, L. Rimell, L. A. Hendricks, S. Legassick, G. Irving, I. Gabriel, Ethical and social risks of harm from Language Models (2021) 64.

[2] P. Delobelle, E. Tokpo, T. Calders, B. Berendt, Measuring Fairness with Biased Rulers: A Comparative Study on Bias Metrics for Pre-trained Language Models, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2022, pp. 1693–1706. doi:10.18653/v1/2022.naacl-main.122.

[3] R. Berk, H. Heidari, S. Jabbari, M. Kearns, A. Roth, Fairness in Criminal Justice Risk Assessments: The State of the Art (2017-05-27). arXiv:1703.09207.

[4] A. Lundgard, Measuring justice in machine learning, arXiv preprint arXiv:2009.10050 (2020).

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[6] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, et al., Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, arXiv preprint arXiv:2206.04615 (2022).

[7] Chatgpt, https://chat.openai.com/, 2023.

[8] D. Levi, Chatgpt crosses 1 million users five days after launch (December, 2022). (accessed: 24. 5. 2023).

[9] A. Salinas, L. Penafiel, R. McCormack, F. Morstatter, "i'm not racist but...": Discovering bias in the internal knowledge of large language models, arXiv preprint arXiv:2310.08780 (2023).

[10] K. Khandelwal, M. Tonneau, A. M. Bean, H. R. Kirk, S. A. Hale, Casteist but not racist? quantifying disparities in large language model bias between india and the west, arXiv preprint arXiv:2309.08573 (2023).

[11] Muslim-violence bias, https://github.com/google/BIG-bench/blob/main/bigbench/benchmark_tasks/muslim_violence_bias/README.md, 2022.

[12] D. Hellman, Discrimination and Social Meaning, in: K. Lippert-Rasmussen (Ed.), The Routledge Handbook of the Ethics of Discrimination, 1 ed., Routledge, 2017-08-23, pp. 97–107. doi:10.4324/9781315681634-10.

[13] A. P. Association, Bias, https://dictionary.apa.org/bias, 2023. (accessed: 25.05.2023).

[14] A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language

corpora contain human-like biases 356 (2017-04-14) 183–186. doi:10.1126/science.aal4230.

[15] B. Friedman, H. Nissenbaum, Bias in computer systems, ACM Transactions on information systems (TOIS) 14 (1996) 330–347.

[16] S. L. Blodgett, S. Barocas, H. Daumé III, H. Wallach, Language (Technology) is Power: A Critical Survey of "Bias" in NLP, 2020-05-29. arXiv:2005.14050.

[17] D. K. Mulligan, J. A. Kroll, N. Kohli, R. Y. Wong, This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology 3 (2019-11-07) 1–36. doi:10.1145/3359221.

[18] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, J. Vertesi, Fairness and abstraction in sociotechnical systems, in: Proceedings of the conference on fairness, accountability, and transparency, 2019, pp. 59–68.

[19] R. Adcock, D. Collier, Measurement validity: A shared standard for qualitative and quantitative research, American political science review 95 (2001) 529–546.

[20] S. Messick, Validity, ETS research report series 1987 (1987) i–208.

[21] D. J. Hand, Measurement: A very short introduction, Oxford University Press, 2016.

[22] A. Z. Jacobs, H. Wallach, Measurement and Fairness, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, ACM, 2021-03-03, pp. 375–385. doi:10.1145/3442188.3445901.

[23] H. Kirk, Y. Jun, H. Iqbal, E. Benussi, F. Volpin, F. A. Dreyer, A. Shtedritski, Y. M. Asano, Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models, 2021. arXiv:2102.04130.

[24] J. Dhamala, T. Sun, V. Kumar, S. Krishna, Y. Pruksachatkun, K.-W. Chang, R. Gupta, BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, ACM, 2021-03-03, pp. 862–872. doi:10.1145/3442188.3445924.

[25] S. Touileb, L. Øvrelid, E. Velldal, Occupational Biases in Norwegian and Multilingual Language Models, in: Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), Association for Computational Linguistics, 2022, pp. 200–211. doi:10.18653/v1/2022.gebnlp-1.21.

[26] S. Alnegheimish, A. Guo, Y. Sun, Using natural sentence prompts for understanding biases in language models, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022, pp. 2824–2830.

[27] M. Bartl, M. Nissim, A. Gatt, Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias, arXiv preprint arXiv:2010.14534 (2020).

[28] D. de Vassimon Manela, D. Errington, T. Fisher, B. van Breugel, P. Minervini, Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021, pp. 2232–2242.

[29] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai, Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings, NeurIPS 2016 (2016) 9.

[30] S. Jentzsch, C. Turan, Gender Bias in BERT - Measuring and Analysing Biases through Sentiment Rating in a Realistic Downstream Classification Task, in: Proceedings of the

4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), Association for Computational Linguistics, 2022, pp. 184–199. doi:10.18653/v1/2022.gebnlp-1.20.

[31] R. Wolfe, A. Caliskan, Low Frequency Names Exhibit Bias and Overfitting in Contextualizing Language Models (2021-10-01). arXiv:2110.00672.

[32] A. Silva, P. Tambwekar, M. Gombolay, Towards a Comprehensive Understanding and Accurate Evaluation of Societal Biases in Pre-Trained Transformers, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2021, pp. 2383–2389. doi:10.18653/v1/2021.naacl-main.189.

[33] O. Deery, K. Bailey, The Bias Dilemma: The Ethics of Algorithmic Bias in Natural-Language Processing 8 (2022).

[34] R. Basu, The Specter of Normative Conflict, in: E. Beeghly, A. Madva (Eds.), An Introduction to Implicit Bias, 1 ed., Routledge, 2020-03-27, pp. 191–210. doi:10.4324/9781315107615-10.

[35] M. Nadeem, A. Bethke, S. Reddy, StereoSet: Measuring stereotypical bias in pretrained language models (2020-04-20). arXiv:2004.09456.

[36] W. B. Gallie, IX.—Essentially Contested Concepts 56 (1956-06-01) 167–198. doi:10.1093/aristotelian/56.1.167.

[37] D. Hellman, MEASURING ALGORITHMIC FAIRNESS 106 (2023).

[38] R. Binns, Fairness in Machine Learning: Lessons from Political Philosophy (2018) 11.

[39] A. Sotnikova, Y. T. Cao, H. Daumé III, R. Rudinger, Analyzing Stereotypes in Generative Text Inference Tasks, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, 2021, pp. 4052–4065. doi:10.18653/v1/2021.findings-acl.355.

[40] C. C. Perez, Invisible women: Data bias in a world designed for men, Abrams, 2019.

[41] K. Crawford, The trouble with bias. keynote at neurips (2017).

[42] S. Barocas, M. Hardt, A. Narayanan, Fairness and machine learning: Limitations and opportunities, MIT Press, 2023.

[43] J.-Y. Yao, K.-P. Ning, Z.-H. Liu, M.-N. Ning, L. Yuan, Llm lies: Hallucinations are not bugs, but features as adversarial examples, arXiv preprint arXiv:2310.01469 (2023).

[44] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).

[45] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, et al., Lamda: Language models for dialog applications, arXiv preprint arXiv:2201.08239 (2022).

[46] S. L. Blodgett, G. Lopez, A. Olteanu, R. Sim, H. Wallach, Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 1004–1015. URL: https://aclanthology.org/2021.acl-long.81. doi:10.18653/v1/2021.acl-long.81.

[47] L. Alexander, M. Moore, Deontological Ethics, in: E. N. Zalta (Ed.), The Stanford Ency-

clopedia of Philosophy, Winter 2021 ed., Metaphysics Research Lab, Stanford University, 2021.

[48] J.-C. Heilinger, H. Kempt, The ethics of (generative) ai. (forthcoming), Critical AI 2 (2024).

[49] L. Lucy, D. Bamman, Gender and Representation Bias in GPT-3 Generated Stories, in: Proceedings of the Third Workshop on Narrative Understanding, Association for Computational Linguistics, 2021, pp. 48–55. doi:10.18653/v1/2021.nuse-1.5.

[50] K. Lippert-Rasmussen, Born free and equal?: A philosophical inquiry into the nature of discrimination, Oxford University Press, 2013.

[51] S. Moreau, What is discrimination?, Phil. & Pub. Aff. 38 (2010) 143.

[52] K. Lippert-Rasmussen, Indirect discrimination is not necessarily unjust, Journal of Practical Ethics 2 (2014).

[53] A. Slavny, T. Parr, Harmless discrimination, Legal Theory 21 (2015) 100–114.

[54] I. Databases, Practice relating to rule 7. the principle of distinction between civilian objects and military objectives section a. the principle of distinction, 2024. URL: https://ihl-databases.icrc.org/en/customary-ihl/v2/rule7.

[55] M. Loi, M. Christen, Choosing how to discriminate: Navigating ethical trade-offs in fair algorithmic design for the insurance sector 34 (2021-12) 967–992. doi:10.1007/s13347-021-00444-9.

[56] G. M. Johnson, Algorithmic bias: On the implicit biases of social technology 198 (2021-10) 9941–9961. doi:10.1007/s11229-020-02696-y.

[57] D. Gala, M. O. Khursheed, H. Lerner, B. O'Connor, M. Iyyer, Analyzing gender bias within narrative tropes, in: Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science, Association for Computational Linguistics, Online, 2020, pp. 212–217. doi:10.18653/v1/2020.nlpcss-1.23.

[58] S. L. Smith, M. Choueiti, A. Prescott, K. Pieper, Gender roles & occupations: A look at character attributes and job-related aspirations in film and television, Geena Davis Institute on Gender in Media (2012) 1–46.

[59] M. Sap, M. C. Prasettio, A. Holtzman, H. Rashkin, Y. Choi, Connotation frames of power and agency in modern films, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2329–2334. doi:10.18653/v1/D17-1247.

[60] E. Fast, T. Vachovsky, M. Bernstein, Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 10, 2016, pp. 112–120.

[61] E. L. Papadaki, Feminist Perspectives on Objectification, in: E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, Spring 2021 ed., Metaphysics Research Lab, Stanford University, 2021.

[62] S. L. Bartky, Femininity and domination: Studies in the phenomenology of oppression, Routledge, 2015.

[63] S. M. Okin, Justice, Gender, and the Family, New York: Basic Books, 1989.

[64] A. Allen, Feminist Perspectives on Power, in: E. N. Zalta, U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy, Fall 2022 ed., Metaphysics Research Lab, Stanford University, 2022.

[65] Europol, Chatgpt - the impact of large language models on law enforcement, a tech watch

flash report from the europol innovation lab (2023).

[66] J. Dempere, K. P. Modugu, A. Hesham, L. Ramasamy, The impact of chatgpt on higher education, Dempere J, Modugu K, Hesham A and Ramasamy LK (2023) The impact of ChatGPT on higher education. Front. Educ 8 (2023) 1206936.

[67] C. Faife, Debt collectors want to use ai chatbots to hustle people for money, https://www.vice.com/en/article/bvjmm5/debt-collectors-want-to-use-ai-chatbots-to-hustle-people-for-money, 18-05-2023. (accessed: 24. 5. 2023).

[68] L. Laestadius, A. Bishop, M. Gonzalez, D. Illenčík, C. Campos-Castillo, Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot replika, New Media & Society (2022) 14614448221142007.

[69] The us military is taking generative ai out for a spin, https://www.bloomberg.com/news/newsletters/2023-07-05/the-us-military-is-taking-generative-ai-out-for-a-spin?ref=hackernoon.com/, 2023.

[70] J.-C. Heilinger, The ethics of ai ethics. a constructive critique, Philosophy & Technology 35 (2022) 61.

[71] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias, in: Ethics of data and analytics, Auerbach Publications, 2022, pp. 254–264.

[72] J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in: Conference on fairness, accountability and transparency, PMLR, 2018, pp. 77–91.

[73] M. Raghavan, S. Barocas, J. Kleinberg, K. Levy, Mitigating bias in algorithmic hiring: Evaluating claims and practices, in: Proceedings of the 2020 conference on fairness, accountability, and transparency, 2020, pp. 469–481.