You Reap What You Sow: On the Challenges of Bias Evaluation Under Multilingual Settings

Anonymous ACL submission

Abstract

Evaluating bias, fairness, and social impact in monolingual language models is a difficult task. This challenge is further compounded when language modeling occurs in a multilingual context. Considering the implication of evaluation biases for large multilingual language models, we situate the discussion of bias evaluation within a wider context of social scientific research with computational work. We highlight three dimensions of developing multilingual bias evaluation frameworks: (1) increasing transparency through documentation, (2) expanding targets of bias beyond gender, and (3) addressing cultural differences that exist between languages. We further discuss the power dynamics and consequences of training large language models and recommend that researchers remain cognizant of the ramifications of developing such technologies.

1 Introduction

Machine learning (ML) systems, especially large language models (LLMs), are prone to (re)produce harmful outcomes and social biases (Bender et al., 2021; Raji et al., 2021; Blodgett et al., 2020; Aguera y Arcas et al., 2018). Despite recent advances in LLMs (Bender and Koller, 2020), they have shown to disproportionately produce harmful content when addressing certain topics (Gehman et al., 2020; Lin et al., 2021) and demographics (Sheng et al., 2019; Liang et al., 2021; Dev et al., 2021a)—in part due to the training data used (Dunn, 2020; Gao et al., 2020; Bender et al., 2021), and the design of modeling processes (Talat et al., 2021; Hovy and Prabhumoye, 2021). In response, previous work has explored ways in which such social biases can be measured and counteracted (Nangia et al., 2020; Gehman et al., 2020; Czarnowska et al., 2021). Typically, these issues have been addressed either by conceptualizing the underlying systemic discrimination as "bias" or by developing evaluation datasets that shed light on how LLMs produce harmful social outcomes. However, in the former case, as Blodgett et al. (2020) points out, these conceptualizations often lack clear descriptions, e.g., type of systemic discrimination and affected demographics. This results in a highly under-specified "bias", which could lead to a downstream issue in the validity of the technical approaches that are developed (Blodgett et al., 2021). Similarly, the ill-defined "bias" is further compounded by the specifics of many benchmarks. Often, benchmarks exhibit discrepancies between understandings of the unobservable theoretical constructs against which "bias" is being measured and their operationalization (Jacobs and Wallach, 2021; Friedler et al., 2021). Furthermore, many prior benchmark datasets were developed with specific modeling architectures in mind (Nangia et al., 2020). They are limited to English and are culturally Anglo-centric.[1]

In this position paper, we present an overview of the current state-of-the-art concerning challenges and measures taken to address bias in language models. Specifically, we document the challenges of evaluating language models, with a focus on the generation of harmful text. By engaging our challenges with the relevant social scientific literature, we propose (1) a more transparent evaluation of bias via scoping and documentation, (2) focusing on the diversity of stereotypes for increased inclusivity, (3) careful curation of culturally aware datasets, and (4) creation of general bias measures that are independent of model architecture but capture the context of the task.

We recognize that many of the challenges that we have encountered and described here are large open problems that will require joint work to address. Our goal is to analyze these challenges and provide scaffolding for future work.

2 Grounding Bias, Fairness and Social Impact across Disciplines

Considering biases in socio-technical systems as a purely technical construct is an insufficient consideration of the problem (Blodgett et al., 2020). In this section, we situate LLMs, and their applications, within the wider interdisciplinary literature on social harms and discrimination.

2.1 Social Discrimination

Issues of socially discriminatory (human and technological) systems have long been the subject of study

[1] For example, the BigScience biomedical working group has estimated that 82% of evaluation datasets in the biomedical and clinical field are for corpora in English (Datta et al., 2021).

for scholars across disciplines, e.g. in Science and Technology Studies (Haraway, 1988), discard studies (Lepawsky, 2019), social anthropology (Douglas, 1978), philosophy of democracy (Fraser, 1990), gender and LGBTQIA+ studies (Spade, 2015; Rajunov and Duane, 2019; Keyes et al., 2021; D'Ignazio and Klein, 2020), media studies (Gitelman, 2013), archival studies (Agostinho et al., 2019), sociolinguistics (Labov, 1986; Cheshire, 2007), and critical race theory (Noble, 2018; Benjamin, 2019).[2]

Scholars argue that technical systems are embedded in social contexts (Lepawsky, 2019; Haraway, 1988) and are therefore necessarily evaluated as socio-technical systems interacting with complex social hierarchies (Winner, 1980; Benjamin, 2019; Costanza-Chock, 2018; Friedler et al., 2021). When technological systems prioritize majorities, there is a risk they oppress minorities at the personal, communal, and institutional levels (Costanza-Chock, 2018). Haraway (1988) argues that researchers default to a "view from nowhere", without reflecting on the context or use of their research. This default view often represents the interests of dominant majorities, disregarding knowledges from marginalized communities. Considering machine learning systems, Chun (2021) argues that the development of such technological systems relies on faulty assumptions (e.g., that past data collections can adequately and fairly predict future human behavior) which can lead to embedded social biases. Situating ourselves in the wider academic literature of social discrimination and marginalization, compels us to recognize that our technical systems must be considered in the social context in which they exist.

2.2 Machine-learned Systems in Social Context

On the topic of socially discriminatory systems within machine learning, Buolamwini and Gebru (2018) and Raji and Buolamwini (2019) show that there are significant disparities along gendered and racialized lines in commercially available facial recognition and analysis systems. Similar issues of discriminatory social biases in natural language processing (NLP) systems have resulted in emerging research dedicated to the identification, quantification (e.g. Rudinger et al., 2018; De-Arteaga et al., 2019; Czarnowska et al., 2021), and mitigation of bias (Bolukbasi et al., 2016; Sun et al., 2019; Garimella et al., 2021) in NLP systems.

However, these methods tend to obscure rather than remove social biases (Gonen and Goldberg, 2019), and are particularly brittle when applied to complex, contextual language representations (Dev et al., 2020).

Further, operationalization of under-specified "bias" has varied widely across studies, and in some cases has been internally inconsistent with their stated goals (Blodgett et al., 2020; Jacobs and Wallach, 2021). The recent surge of LLMs is no exception to such concerns. Hovy and Prabhumoye (2021); Talat et al. (2021), and Cao and Daumé III (2020) argue that socially discriminatory biases can be encoded in several stages of the LLM development process (Biderman and Scheirer, 2020), including data sampling, annotation, selection of input representations or model, research design, and how the models are situated with regards to the language communities that they are applied to. Language generation models, despite their inference-time flexibility, are particularly susceptible to reproducing hegemonic social biases and generating offensive language, even when not explicitly prompted to do so (Sheng et al., 2021; Wallace et al., 2019; Bender et al., 2021).

In efforts to address the expression of such social biases, a number of bias evaluation benchmarks have been proposed (Dev et al., 2021b; Zhao et al., 2018; Cao and Daumé III, 2020). However, common evaluation benchmarks are fraught with pitfalls in their conceptualization of bias, stereotypes, and harms, including meaningless or poorly formed stereotype constructions, non-intersectional examples, contexts that don't reflect downstream use, and reliance on specific model architectures (Blodgett et al., 2021; Jin et al., 2021). Furthermore, bias evaluation benchmarks often make strong assumptions about the validity, reliability, and existence of observable properties, e.g. pronouns, as signals for unobservable theoretical constructs such as gender (Jacobs and Wallach, 2021). This is particularly problematic when building benchmarks for biases against communities that resist categorization based on observable characteristics (e.g. LGBTQIA+ and racialized people) and leads to reliance on existing stereotypes (Tomasev et al., 2021; Dev et al., 2021a).

This rapid development of NLP resources and tools have further yielded a non-inclusive environment, skewed heavily towards English and Anglo-centric biases (Joshi et al., 2020). Sambasivan et al. (2021) and Chan et al. (2021) contend there remains a significant gap between the communities governing and governed by AI, and advocate for a redistribution of powers and responsibilities in developing responsible AI.

Considering gender bias, Stanczak and Augenstein (2021) show that existing methods (1) largely avoid ethical considerations or evaluations of gender bias, (2) focus primarily on binary gender treatment, in mostly Anglo-centric settings, and (3) employ limited or flawed evaluation methodologies. Such issues are in part exacerbated by the general poverty of documentation of datasets (Gebru et al., 2018; Bender and Friedman, 2018) and machine learning models (Mitchell et al., 2019). One way to mitigate these biases includes creating diverse teams with varied backgrounds and life experiences to assure the expression of diverse perspectives (Monteiro and Castillo, 2019; Nekoto et al., 2020). However, as critiqued by Talat et al. (2021); West et al. (2019), incorporating the diversity factor may be inadequate. Biases in language representations and task

[2]Many recent works on socially biased technological systems are interdisciplinary, e.g., 'Race After Technology: The New Jim Code' (Benjamin, 2019) spans critical race theory, science and technology, Black feminism, and media studies.

models can not only reflect, but also amplify bias present in the datasets (Barocas and Selbst, 2016; Wang et al., 2019). These biases have been investigated and attempts made at creating interpretable representations and providing post-hoc explanations of model predictions.

2.3 Bias, Fairness, and Explainability

Given the grave consequences that inherent or conceptualized biases in ML systems can inflict, *responsible AI* has received a growing amount of research attention (Amershi et al., 2020). Responsible AI refers to the creation of ethical principles for AI and the development of AI systems based on these principles (Dignum, 2017; Schiff, 2020). Colloquially, responsible AI encompasses distinct machine learning fields such as fairness, explainability, privacy, and interpretability. Concretely, how can responsible AI principles best contribute to the development of equitable systems?

Examining this question, Friedler et al. (2021) propose that building just ML systems requires an *a priori* definition of fairness. However, contemporary decision-making systems build on a so-called what-you-see-is-what-you-get (WYSIWYG) approach that implicitly imbibes multiple fairness definitions or world views, leading to a system based on the conflict between the underlying value systems. To tackle this issue, ML engineers should explicitly state the underlying systemic values, as systems will inevitably comprise certain assumptions (Birhane et al., 2021). Thus, implying that biases as inherent to these decision-making systems and should be clearly articulated (Bender et al., 2021) by explaining the whys and whats (explainability).

However, a more promising course of action for researchers would be to prioritize fairness in the entire life cycle of a language model. The tendency to consider and mitigate undesirable biases in models after training has completed leaves harmful residues that affect the communities we seek to protect (Dev et al., 2021a). Hence, a fruitful approach could be to reduce systemic unfairness by grounding the discussion on clear definitions of fairness based on input from the communities that could be harmed by the system (Liao and Muller, 2019), explaining the inherent biases, and, if possible, minimizing bias issues by employing the measures discussed in, both, the previous and the following sections.

3 Challenges of Bias

Evaluating the social impacts and harmful biases LLMs exhibit is an important development step. However, despite the increased interest in developing bias benchmarks, the field still faces various challenges in evaluating LLMs with *off-the-shelf* benchmarks. In this section, we provide examples of existing bias measures currently used in NLP. We then discuss the challenges that originate from these: (1) they rely on vague definitions of bias, (2) are restricted to particular model architectures, (3) have limited relevance for different cultural contexts, and (4) are difficult to validate and interpret.

3.1 Examples of Bias Measure Studies

Recently, researchers and practitioners have begun to pay more attention to bias measures in NLP systems (Blodgett et al., 2020; Dev et al., 2021b). One line of work has focused on identifying bias in word embeddings: The Word Embedding Association Test (WEAT, Caliskan et al., 2017) measures bias by comparing the relative distances of two sets of target words (e.g. occupation words: *nurse, doctor*) with respect to two sets of attribute words (e.g., gender attributes: *male, female*)—and has inspired other similar approaches (Kurita et al., 2019; May et al., 2019; Dev et al., 2020).

Although word embeddings may help identify biases in the context of LLMs, it is often difficult to access the learned contextual language representations of the model (Abid et al., 2021; Dev et al., 2020). Furthermore, such methods are developed to address static word embeddings rather than the dynamic contextual word embeddings LLMs rely on (Subramonian, 2021).

Another research direction is the use of causal inference for measuring biases in LLMs, for example to analyze if the generated text by an LLM is affected considerably by only changing the protected attributes or categories in the input (Huang et al., 2020; Madaan et al., 2021; Cheng et al., 2021). In line with this idea, Huang et al. (2020) used a sentiment classifier to quantify and reduce the sentiment bias existent in LLMs. Similarly, the CrowS-Pairs benchmark (Nangia et al., 2020) leverages the paradigm of minimal pairs to contrast sentences expressing stereotypes against social categories with the same sentences addressing different social categories. Crows-Pairs is designed such for language models to be probed for disparate behavior between the sentences pairs, with the hypothesis that systematic difference in the treatment reflecting the preference for stereotype indicates the presence of bias in the language models. Other examples of bias measures benchmarks include StereoSet (Nadeem et al., 2020), WinoMT (Stanovsky et al., 2019), BBQ (Parrish et al., 2021), BOLD (Dhamala et al., 2021), and Toxicity Comment Classification competition (Jigsaw, 2017).

3.2 Defining Bias

The term "bias" is overloaded in the ML and NLP communities, as it is used in both the lay (a prejudice towards or against some entity) and the statistical sense (a systematic deviation from a distribution's mean) (Campolo et al., 2018). Moreover, researchers often refer to vague definitions of bias and gloss over the details, which results in methods that lack specificity (Blodgett et al., 2020). When discussing methods to address bias, it is critical to be precise about the bias being addressed.

Bias can, for instance, be made more specific by being defined along socially relevant dimensions. Nangia et al. (2020) consider the protected categories from the US Equal Employment Opportunities Commission and Queer in AI[3] uses a similar list (*gender identity and*

[3]Queer in AI (`http://queerinai.org/`) is a grass-

expression, sexual orientation, disability, neurodivergence, skill set, physical appearance, body size, race, caste, age, nationality, citizenship status, colonial experience, religion), yet other characteristics may be relevant elsewhere in the world (e.g. illness, migrant, and social status). However, protected classes are only one dimension along which to define bias; researchers should also be mindful of political biases and biases resulting from the focus on prestigious, highly resourced language varieties, in additions to the intersections of multiple dimensions (Kearns et al., 2018; Buolamwini and Gebru, 2018; Crenshaw, 1991).

With respect to any of the aforementioned dimensions, a "bias" is a preferential disposition towards or against an entity. Colloquially, it is perceived negatively and considered to be unfair treatment. As pointed out by Barocas et al. (2017), biases in language models can manifest in the form of *quality-of-service* and *representation* disparities. A quality-of-service bias describes subpar performance of a language model when used by a particular group. For example, LLM-driven machine translation systems provide significantly better support for "prestigious", high-resource languages, and consequently deny quality performance to individuals who do not speak these languages (Nekoto et al., 2020). Furthermore, in fundamental NLP tasks such as coreference resolution, LLMs can fail for people who use neopronouns, and often capture meaningless representations for language associated with trans and non-binary individuals. (Cao and Daumé III, 2020; Dev et al., 2021a). Additionally, Blodgett et al. (2018) show that parsing systems trained primarily on White Mainstream American English exhibit disparate performance on African American English and Tan et al. (2020) show that English question answering and machine translation systems often fail on the morphological variation that is often present in non-prestige and Learner Englishes.

Representation biases consist of stereotypes and under-representation (or over-representation) of data or model outputs. Stereotyping is a cognitive process that manifests from often negative cultural norms about a characteristic; stereotyping permeates what people do, say, or write. A long line of work has shown that language models capture social stereotypes, for example, with respect to binary gender and occupations (Zhao et al., 2018; Bordia and Bowman, 2019; de Vassimon Manela et al., 2021). With regard to (under)representation, in MIMIC-III, a clinical notes dataset, only 1.9% of patients identify as Asian, in comparison to 71.5% who identify as white (Chen et al., 2020). Furthermore, blocklists in the Colossal Clean Crawled Corpus (C4) dataset disproportionately filter words related to queerness and language that is not White-aligned English (Dodge et al., 2021). Notably, quality-of-service and representation biases are not mutually exclusive; for instance, the brittle representations learned by a LLM for language associated with trans and non-binary individuals largely stems from the severe under-representation of this in training data (Dev et al., 2021a; Barocas and Selbst, 2016).

The breakdown of biases into quality-of-service and representation disparities is only one of many possible lenses. It is further critical to explicitly consider biases stemming from disparities in resources, broadly defined in terms of data availability, time to invest into dataset curation, access to compute resources, financial resources, and more (Bender et al., 2021).

3.3 Overreliance on Model Architectures

Current benchmarks often measure bias in specific downstream tasks (e.g. Machine Translation (Stanovsky et al., 2019), Question Answering (Parrish et al., 2021), or Text Generation (Dhamala et al., 2021)), while others focus on bias in LLMs more generally (e.g Kurita et al., 2019; Nadeem et al., 2020; Nangia et al., 2020). This has the advantage of being more widely applicable, as many NLP systems are based on LLMs, and it avoids the need for creating and validating a new benchmark for each possible downstream task. Yet, when the benchmarks heavily rely on the model architecture rather than the task specification, quantitative comparison between different models based on these benchmarks is no longer possible. In such cases, it also becomes more difficult to assess the validity of the bias measure in how it relates to other benchmarks (criterion validity) and the more abstract notion of fairness (construct validity).

Some researchers circumvent this problem by adapting the original bias metric, but care should be taken when doing so. For instance, bias metrics originally developed for masked language models have been adapted by using perplexity (e.g. Nadeem et al., 2020) or prompting (e.g. Gao et al., 2021; Sanh et al., 2021) instead. While these could still result in important insights, they also open new questions. Are the underlying assumptions of the bias measure still valid? Can you compare the bias metrics across different (future) types of models? Do the results of the initial validation of the benchmark still hold? And how does the kind of training data impact the evaluation that assumes a different training domain (e.g., legal texts vs. social media)?

While bias is ideally defined independently of the particular model architecture—not least because implementations change over time—we should not fall into a generalization trap either. As argued before, bias is inherent to systems and context-sensitive, and we should not strive for a panacea bias measure. Instead, the goal should be to develop methods that are task-specific yet independent of a given architecture, to the degree that this is possible. Researchers should keep this tension between task- and architecture-specific measures in mind when designing methods for measuring biases in LLMs.

roots D&I organization that seeks to empower queer and trans researchers in AI and advance research at the intersections of AI and queerness. Their list of categories can be found here: http://queerinai.org/code-of-conduct.

3.4 Bias Measures are Anglo-centric

Despite the need for evaluating LLMs for a wide range of languages, bias benchmarks that cover non-English languages are rare (Zhou et al., 2019; Joshi et al., 2020). As a solution, simply translating existing English benchmarks is not ideal: manual translation is a labor-intensive and highly skilled task, while automated translations are prone to errors and could potentially introduce new algorithmic sources of bias. Moreover, translated benchmarks may only test for Anglo-centric biases, which do not necessarily hold in many non-Western cultural contexts. For instance, many gender bias evaluations focus on Western professions, which are grammatically gendered in some languages (Chen et al., 2021; Zhou et al., 2019) or may not cover other prevalent occupations outside the U.S. (Escudé Font and Costa-jussà, 2019). WinoMT (Stanovsky et al., 2019) is one of the few benchmarks that covers multiple languages, but it comes with its own downsides. The sentences are generated from templates that capture a limited range of actual language use; the samples are translated from English examples, which may not reflect how stereotypes would occur in other languages; and the scope is limited to machine translation systems, and therefore WinoMT may not be suitable for multilingual models that are not trained on this specific task. The tightly coupled nature of bias and cultural context should be emphasized when designing a multilingual bias benchmark.

3.5 Validity of Bias Measures

Towards making NLP systems more just, we must understand the flaws of common bias measures and develop better guidelines to address biases. According to Jacobs and Wallach (2021) and Blodgett et al. (2021), bias measures are measurement models which link observable properties, e.g., quality-of-service and representational biases, with unobservable theoretical constructs such as social discrimination, power dynamics, and systemic oppression. Consequently, bias measures are deeply political. Notably, a vast majority of bias measures themselves rely on other measurement models, such as the presence of gendered pronouns, to infer theoretical protected categories, e.g., gender. Moreover, bias measures may cause further epistemic violence onto the marginalized by creating a veneer of fairness, in spite of ongoing marginalization (Gonen and Goldberg, 2019; Talat et al., 2021; Jacobs and Wallach, 2021). In ensuring the reliability, validity, and correct interpretation of bias measures, it is critical to examine all components in a bias measurement method.

Upstream measurement models that infer protected categories can be unreliable or even non-existent. For instance, pronouns and gendered names are usually employed as proxies for binary gender, which is problematic (Dev et al., 2021a). Furthermore, characteristics like sexuality and disability are usually unobservable, which can lead to a reliance on hegemonic stereotypes and unnatural language in bias evaluation benchmarks (Tomasev et al., 2021; Hutchinson et al., 2020).

With regard to validity, Blodgett et al. (2021) reviews how bias measures often rely on operationalization of stereotypes that are invalid for reasons such as misalignment and conflation. Additionally, the mathematical formalization of most bias measures is based on notions of parity-based fairness and do not reflect other conceptualizations of fairness such as distributive justice (Jacobs and Wallach, 2021). Another source of invalidity of bias measures lies in the purported generality of associated benchmarks. Raji et al. (2021) argue that the "instantiation [of benchmarks] in particular data, metrics and practice" undermines the validity of their construction to have "general applicability." Moreover, measurement models for protected categories fallaciously assume that the identities being indirectly observed can be discretized. Hence, Dev et al. (2021b) advocate for documenting the limitations of bias measures and related data in terms of their validity. In this process, it is critical to describe the relationship between the context of the data, model usage, and bias measure at stake.

4 The Elephant in the Room: Power, Privilege, and Point of View

Throughout the paper thus far, we have primarily discussed bias in language models as a mechanical phenomenon. However, it is important to situate these discussions within the context and power dynamics of the way that NLP is practiced — both in research and in application (Miceli et al., 2022). In this section, we discuss sociopolitical influences on AI ethics and bias research in NLP. We argue that contemporary developments of LLMs have been an exercise in financial, institutional, ecological, linguistic, and cultural privilege. They are the consequence of the political will to create totalizing technologies and evaluation of bias, fairness and social impact should be viewed as a countervailing power mechanism, although in some cases serve to obscure these.

4.1 Large Language Models are Expensive

The current dominant paradigm in natural language processing is driven by the creation of ever-larger pretrained transformer models (Brown et al., 2020). As the size of LLMs increases, so do the requirements for hardware, energy, and time. For example, GPT-NeoX 20B (Black et al., 2022) was trained for 1830 hours on 96 A100 GPUs, consuming 43.92 MWh of electricity and emitting 23 metric tons of CO_2. Based on the current price listing of the cloud provider the model was trained on, training such a model would cost between 250,000 and 500,000 USD[4]. While this is not on the scale of the largest research programs, it is a significant amount of

[4]The lower end of this range reflects the common practice of giving discounts of up to 50% for large purchases, while the upper end reflects the sticker price of the systems.

money and beyond the funding of many institutions, or beyond their political will to spend.

While the development of such models can contribute towards improving the ability of people with less resources to pursue cutting edge *downstream* research, such pursuits have significant costs and barriers to entry for *upstream* research. This creates a stratification of research, wherein money is a barrier of entry for some forms of research but not for others.

4.2 Language is Multicultural, Language Models are Not

Although there are thousands of spoken languages in the world, the overwhelming majority of LLMs are monolingual and encode white respectability politics (Thylstrup and Talat, 2020; Kerrison et al., 2018) onto minoritized variants of English (Gehman et al., 2020). In this way, the cost of the developing LLMs extends from externalizing computational and infrastructural costs, to externalizing languages and language variants (Lau, 2021). Specifically, the vast majority of LLMs are trained to operate on an unspecified variant of "English" (Bender, 2019), and in some cases Chinese (see Table 1 for a detailed overview of the top 25 LLMs). The dominance of English, and to a lesser degree Chinese, reifies cultural hegemonies and precipitates technological imperialism. Even when researchers seek to include other languages, these purportedly multilingual models often underserve certain languages and communities (Kerrison et al., 2018; Virtanen et al., 2019; Kreutzer et al., 2022; Gururangan et al., 2022). We also note that few of these models have been assessed for bias or fairness (see table 1).

This act relies on two foundations. First, LLMs should only be used for languages that they have been developed for, with the cultural stereotypes that they have been trained on, thus limiting LLMs to be used within a small set of cultural contexts, or casting cultural contexts for which they are trained onto ones that they are not developed for. Second, should a multilingual LLM be trained, its primary data sources will still be in English, whereas the remaining languages will only be incidental to it. Such cultural imperialism is evident from the fact that only 2 of the 14 organizations involved in developing LLMs have teams in multiple countries (see table 1). Further, all multinational LLM efforts, except for one, draw their membership from the USA, UK, Germany, & Australia. GPT-NeoX 20B (Black et al., 2022) is an exception, as it also includes authors from India. A commonly-used resource for developing LLMs, CommonCrawl, relies on data that primarily stems from the US (Dodge et al., 2021) and is written in privileged dialects of English (Dunn, 2020). This prioritization is reflected by 16 teams being physically located in the U.S. Consequently, the current state of LLM development is a totalizing endeavor (Talat et al., 2021), which engages in externalization across a number of axes, as is apparent from the infrastructural and development practices and the efforts to evaluate and mitigate social harms that arise from such technologies.

4.3 Large Language Models Allow Powerful Actors to Control NLP Research

Due to the costs involved with training large language models and the small number of actors who have decided to train them, the overwhelming majority of research studying their properties is not carried out by people who train LLMs. When the actors that do possess the models choose to not publicly release them, which gives the model trainers a level of control over the research that can be conducted with and by these models. Famously, OpenAI's initial announcement of GPT-3 asserted that access to the model would be heavily restricted while the company continued to research ethical interventions in their model. OpenAI is not alone in this; the idea that it is inherently dangerous to release models to the public has been put forth by several other actors in this space (Weidinger et al., 2021a; Askell et al., 2021).

It is essential to recognize that the decisions regarding access and the kind of research that can be conducted on large language models (or any ML models, for that matter) is an inherently political one (Leahy and Biderman, 2021). Regardless of the truth of the aforementioned claims, they are highly contentious political claims and should be treated as such rather than passively accepted.

Direct access to LLMs is important to perform independent research on their datasets, functions, and societal impact (Kandpal et al., 2022; Carlini et al., 2022). While language models produced by the academic research community are widely available for critical examination, commercial systems are often only available through APIs provided by the developers (see table 1 for an overview on access for the 25 largest pretrained language models. Such restrictions to access to the models and resources that they are developed for provide a significant barrier to a) principles of open science andb) research on how the datasets and language models themselves embed and amplify social biases.

5 Addressing Bias

Researchers have developed various strategies to address bias in large language models. As discussed in earlier sections, however, these strategies are insufficient to tackle multiple dimensions of bias. Below, we enumerate a few ways in which bias can be addressed by the research community to effectively engage with our aforementioned concerns: (1) moving towards a more transparent way of evaluating bias, (2) focusing on the diversity of stereotypes and increasing inclusivity, and (3) considering the impact of linguistic and cultural differences on the identification and mitigation of bias in designing culturally comparable datasets. We would like to highlight that these suggestions are not exhaustive. They will, however, guide the work in this area.

5.1 Transparency Through Documentation

Stereotypes and biases cover a broad definition and vary in conceptualization across geographical and cultural contexts. To ensure that the nuances are well communicated and that practitioners understand the applicability of the evaluation approach, we suggest documenting a thorough analysis of the scope. Below, we provide a starting point based on Mitchell et al. (2019); Gebru et al. (2018); Dev et al. (2021b); Blodgett et al. (2020) .

Defining the scope of the approach Blodgett et al. (2020) found that works around bias "often fail to explain what kinds of system behaviors are harmful, in what ways, to whom, and why." It thus becomes imperative to question what underrepresented groups would benefit more from a given evaluation benchmark. We therefore urge researchers and practitioners to clearly specify the demographic a particular method is relevant for. Moreover, given how social hierarchies intertwine tightly with language and may present themselves through its peculiarities, we also encourage researchers to specify the limitations and scope of their approaches.

As an example, we consider the gender bias evaluation in English (Zhao et al., 2018; Stanovsky et al., 2019; Levy et al., 2021; Sharma et al., 2021), where the bias might present itself through strong associations between grammatical constructs like pronouns. The same does not hold true for genderless languages, despite the existence of the bias (Zmigrod et al., 2019). Thus, evaluation benchmarks and approaches do not always transfer well to other languages. Additionally, while such benchmarks use gender associations to professions for their evaluation, this method covers only one aspect of the social hierarchy, and does not address gender bias in language in its entirety. By being binary in nature and tightly coupled to Anglo-centric contexts (see §3) benchmarks are limited in their scope and relevance. While most recent works do include ethical considerations, the limitations and scope are only vaguely specified. We advocate for such limitations to be highlighted and pointed out for the community to have a clearer picture about the steps that need to be taken towards greater inclusivity.

Documenting the demographics Previous work has highlighted the importance of engaging with individuals on the receiving end of the bias (Bender et al., 2021). It thus becomes important to understand the demographics of those involved in the creation of the benchmarks. As previously shown (Al Kuwatly et al., 2020) there exists a relation between annotators' identities and toxicity/bias in dataset. On this basis, we urge the researchers to collect and document the demographic information and annotator attitude scores (Sap et al., 2021). Building upon the same, we encourage the collection and reporting of this information about the researchers involved.

5.2 Diversity Beyond Gender Bias

The majority of previous work on bias has focused particularly on gender bias (Zhao et al., 2018; Stanovsky et al., 2019; Levy et al., 2021; Sharma et al., 2021) and the very few works (Nadeem et al., 2020; Nangia et al., 2020) that take other dimensions of biases into account, have their own shortcomings, as discussed in Section 3. It thus becomes important to diversify the range of bias and stereotypes that are being investigated by research, and covered by a certain evaluation technique. In extending the coverage to more dimensions, context stands as an important aspect of bias. The contextual aspects of bias as represented in language, culture, and history hold a significant role in forming and assessing the bias itself. Hence, as a practice, we encourage researchers to consider these three aspects when constructing bias measures and datasets.

In discussing bias, it is important to note that discrimination does not occur in a vacuum. An act of discrimination against a person may be directed towards several intersecting identities. Considering bias using a single-axis framework makes it impossible to engage with and evaluate the harms extended to the social groups that lie at the intersection of multiple identities (Crenshaw, 1991). In an Indian context, for example, even those who identify as belonging to the "same" caste (Malik et al., 2021), can have varied lived experiences based on class, gender, and other identities. More precisely, it is impossible to disentangle which specific identity a discriminatory act is directed against. Previous works have highlighted the importance of studying intersectional bias (Bender et al., 2021; Buolamwini and Gebru, 2018; Field et al., 2021; Guo et al., 2019; Crenshaw, 1991) but little research has been conducted around addressing such biases (Magee et al., 2021; Guo and Caliskan, 2021). We thus encourage researchers to develop measures and benchmarks which are grounded in intersectional understanding of bias and adequately address the lived experiences of various social groups, towards increased inclusivity and fairness.

Not only can the dimensions and context influence our definitions and approaches to bias, but the categories (values) assigned to each dimension (e.g., age) can also limit our understanding and solution of bias. For instance, the majority of gender-bias evaluation datasets solely deal with binary gender, i.e., male and female, with just a handful covering non-binary genders with only minimal representation (Dev et al., 2021a; Cao and Daumé III, 2020). As a result, category inclusiveness is critical in the development of a high-quality bias evaluation dataset. A set of categories that can act as a starting point are provided by Queer in AI in Section 3.2.

5.3 Acknowledging Differences

Stereotype and bias formation is influenced by culture. As a result, what might be a stereotype in a given culture might not stand relevant in another. For instance, the characterization that `parental leave is for mothers` is considered stereotypical in the United States, but not in Sweden, where parental leave is split between both parents.

	Organizaton	Author Location	Language	Parameters	Model Access	Bias Eval
MT-NLG	Microsoft, NVIDIA	USA	English	530 B	Closed	Smith et al. (2022)
Gopher	DeepMind	USA	English	280 B	Closed	Weidinger et al. (2021b)
ERNIE 3.0	Baidu	China	English, Chinese	260 B	Closed	—
Yuan 1.0	Inspur AI	China	Chinese	245 B	Closed	—
HyperCLOVA	NAVER	Korea	Korean	204 B	Closed	—
PanGu-α	Huawei	China	Chinese	200 B	Closed	—
Jurassic-1	AI21 Labs	Israel	English	178 B	Commercial	—
GPT-3	OpenAI	USA	English	175 B	Commercial	Brown et al. (2020)
LaMDA	Google	USA	English	137 B	Closed	Thoppilan et al. (2022)
Anthropic LM	Anthropic	USA	English	52 B	Closed	Askell et al. (2021)
GPT-NeoX-20B	EleutherAI	Multinational	English	20 B	Open	(Gao et al., 2020; Biderman et al., 2022)
Turing NLG	Microsoft	USA	English	17.2 B	Closed	—
FairSeq Dense	Meta AI	Multinational	English	13.0 B	Open	—
mT5	Google	USA	Multilingual	13.0 B	Open	???
ByT5	Google	USA	English	13.0 B	Open	—
T5	Google	USA	English	11.0 B	Open	???
CPM 2.1	Tsinghua University	China	Chinese	11.0 B	Open	—
Megatron 11B	NVIDIA	USA	English	11.0 B	Open	???
WuDao-GLM-XXL	Beijing Academy of AI	China	Chinese	10.0 B	Open	???
WuDao-GLM-XXL	Beijing Academy of AI	China	English	10.0 B	Open	???
BlenderBot	Meta AI	USA	English	9.4 B	Open	no
Megatron-LM	NVIDIA	USA	English	8.3 B	Closed	—
XGLM	Meta AI	Multinational	Multilingual	7.5 B	Open	—
GPT-J-6B	EleutherAI	Multinational	English	6.0 B	Open	(Gao et al., 2020; Biderman et al., 2022)

Table 1: The 25 largest pretrained dense language models, ranging from 6 billion parameters to 530 billion. Models are overwhelmingly trained by teams located in the US and on English text. Less than half of the language models were evaluated for bias by their creators.

Previous sections have criticized the Anglo-centricity in the research of NLP bias and the influence on languages other than English. In particular, the lack of culturally-aware datasets limits the degree to which future NLP algorithms can be evaluated for biases. More crucially, these unspecified languages and cultures are on the receiving end of unmanaged effects. As a result, researchers are encouraged to develop bias datasets and benchmarks for non Anglo-centric cultures and languages (Bender et al., 2021). Involving experts in related areas, especially participants with lived experiences of language-related harms, might aid decisions at all parts of this process, e.g. deciding what groups and content to include in research or dataset design (Liao and Muller, 2019; Dev et al., 2021a; McMillan-Major et al., 2022). Overall, having culturally diverse and comparable datasets for a diverse set of languages (ideally covering all languages) is critical for evaluating multilingual models. Moreover, the applicability of bias measures across various languages suggests the necessity for cross-linguistic metrics or measurements that can be extended to different languages or cultures (Zhou et al., 2019; Escudé Font and Costa-jussà, 2019; Malik et al., 2021).

6 Conclusion

Recent improvements in LLMs to mimic human text have led to a surge in research that seeks to identify and address the harms arising from their training and deployment. However, the considerations on social harms that arise has been limited to narrow, Anglocentric, contradictory, and often underspecified definitions of fairness and bias. Furthermore, the development of contemporary methods has conflated task-specific and architecture-specific designations. Compounded with the structural inequalities around resources, language, and identity, this has yielded an overreliance on prestige forms of English for developing LLMs and interrogating and addressing the social biases that they harbor. Situating these methods within such Englishes has had the consequence of over-emphasizing Western-centric social categories. Moreover, datasets for evaluating social biases in LLMs have traditionally failed to denote and specify the context within which biases are situated. Such concerns have been the cause for questions around the validity of the developed measures, and in particular for multilingual LLMs.

To address such challenges, we propose that developing methods for multilingual LLMs requires researchers to provide thorough documentation of their approaches, including documenting the scope, demographics of speakers, and potential annotators. Additionally, we also recommend that researchers situate their bias evaluation methods within the specific context of the languages that the model operates on. In doing so, bias evaluation methods can be made to specifically address biases under the conditions and contexts that they occur in each of the model's languages. Furthermore, we recommend that researchers examine diversity issues beyond gender bias, with a particular focus on intersectional issues (Guo and Caliskan, 2021).

Finally, we recommend that researchers are cognizant of the social and environmental harms that developing LLMs have. For instance, developing ever-larger language models that achieve marginal improvements for English may bring a smaller benefit than developing a LLM for other languages. Thus, in a consideration of developing a new language model, we implore researchers to consider ways in which harms can be limited, or the benefits can come to compensate for their costs.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent Anti-Muslim Bias in Large Language Models. In *AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*.

Daniela Agostinho, Catherine D'Ignazio, Annie Ring, Nanna Bonde Thylstrup, and Kristin Veel. 2019. Uncertain Archives: Approaching the Unknowns, Errors, and Vulnerabilities of Big Data through Cultural Theories of the Archive. *Surveillance & Society*, 17(3/4):422–441.

Blaise Aguera y Arcas, Alexander Todorov, and Margaret Mitchell. 2018. Do algorithms reveal sexual orientation or just expose our stereotypes?

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.

Saleema Amershi, Ece Kamar, Kristin Lauter, Jenn Wortman Vaughan, and Hanna Wallach. 2020. Research Supporting Responsible AI.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: from allocative to representational harms in machine learning. special interest group for computing. *Information and Society (SIGCIS)*, 2.

Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review*, 104(3).

Emily Bender. 2019. The #BenderRule: On Naming the Languages We Study and Why It Matters.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Ruha Benjamin. 2019. *Race after technology: abolitionist tools for the new Jim code*. Polity, Medford, MA.

Stella Biderman, Kieran Bicheno, and Leo Gao. 2022. Datasheet for the Pile. *arXiv:2201.07311 [cs]*. ArXiv: 2201.07311.

Stella Biderman and Walter Scheirer. 2020. Pitfalls in machine learning research: Reexamining the development cycle. In *"I Can't Believe It's Not Better!" NeurIPS 2020 workshop*.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2021. The Values Encoded in Machine Learning Research. *arXiv:2106.15590 [cs]*. ArXiv: 2106.15590.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An Open-Source Autoregressive Language Model.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In *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)*, pages 1004–1015, Online. Association for Computational Linguistics.

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. 2018. Twitter Universal Dependency parsing for African-American and mainstream American English. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Melbourne, Australia. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and Reducing Gender Bias in Word-Level Language Models. In *Proceedings of the 2019 Conference of the North*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA. PMLR.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334).

Alex Campolo, Madelyn Sanfilippo, Meredith Whittaker, and Kate Crawford. 2018. AI now 2017 report. In *AI now 2017 symposium and workshop*. AI Now Institute at New York University. Edition: AI Now 2017 Symposium and Workshop.

Yang Trista Cao and Hal Daumé III. 2020. Toward Gender-Inclusive Coreference Resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*.

Alan Chan, Chinasa T. Okolo, Zachary Terner, and Angelina Wang. 2021. The Limits of Global Inclusion in AI Development. *arXiv:2102.01265 [cs]*. ArXiv: 2102.01265.

John Chen, Ian Berlot-Attwell, Xindi Wang, Safwan Hossain, and Frank Rudzicz. 2020. Exploring text specific and blackbox fairness algorithms in multimodal clinical NLP. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 301–312, Online. Association for Computational Linguistics.

Yan Chen, Christopher Mahoney, Isabella Grasso, Esma Wali, Abigail Matthews, Thomas Middleton, Mariama Njie, and Jeanna Matthews. 2021. Gender Bias and Under-Representation in Natural Language Processing Across Human Languages. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 24–34, Virtual Event USA. ACM.

Lu Cheng, Ahmadreza Mosallanezhad, Paras Sheth, and Huan Liu. 2021. Causal Learning for Socially Responsible AI. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*.

Jenny Cheshire. 2007. An untitled review of "style and sociolinguistic variation". *Language*, 83(2):432–435.

Wendy Hui Kyong Chun. 2021. *Discriminating data: correlation, neighborhoods, and the new politics of recognition*. The MIT Press, Cambridge, Massachusetts.

Sasha Costanza-Chock. 2018. Design Justice, A.I., and Escape from the Matrix of Domination. *Journal of Design and Science*.

Kimberle Crenshaw. 1991. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43(6):1241–1299.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Debajyoti Datta, Jason A. Fries, Michael McKenna, Aurélie Névéol, Vassilina Nikoulina, and Maya Varma. 2021. Challenges in language modelling for biomedicine.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128, Atlanta GA USA. ACM.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. 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*, pages 2232–2242, Online. Association for Computational Linguistics.

Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7659–7666.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021a. Harms of gender exclusivity and challenges in non-binary representation in language technologies.

In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Nanyun Peng, and Kai-Wei Chang. 2021b. What do bias measures measure?

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 862–872, Virtual Event Canada. ACM.

Catherine D'Ignazio and Lauren F. Klein. 2020. *Data feminism*. Strong ideas series. The MIT Press, Cambridge, Massachusetts.

Virginia Dignum. 2017. Responsible artificial intelligence: Designing ai for human values. *ICT Discoveries*.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mary Douglas. 1978. *Purity and danger: an analysis of the concepts of pollution and taboo*, repr edition. Routledge, London. OCLC: 248038797.

Jonathan Dunn. 2020. Mapping languages: the Corpus of Global Language Use. *Language Resources and Evaluation*, 54(4):999–1018.

Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy. Association for Computational Linguistics.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A Survey of Race, Racism, and Anti-Racism in NLP. In *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)*, pages 1905–1925, Online. Association for Computational Linguistics.

Nancy Fraser. 1990. Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy. *Social Text*, (25/26):56.

Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (Im)possibility of fairness: different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64(4):136–143.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv:2101.00027 [cs]*. ArXiv: 2101.00027.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *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)*, pages 3816–3830, Online. Association for Computational Linguistics.

Aparna Garimella, Akhash Amarnath, Kiran Kumar, Akash Pramod Yalla, Anandhavelu N, Niyati Chhaya, and Balaji Vasan Srinivasan. 2021. He is very intelligent, she is very beautiful? On Mitigating Social Biases in Language Modelling and Generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4534–4545, Online. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for Datasets. *arXiv:1803.09010 [cs]*. ArXiv: 1803.09010.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Lisa Gitelman. 2013. *Raw data is an oxymoron*. MIT press.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of the 2019 Conference of the North*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Anhong Guo, Ece Kamar, Jennifer Wortman Vaughan, Hanna Wallach, and Meredith Ringel Morris. 2019. Toward fairness in ai for people with disabilities: A research roadmap.

Wei Guo and Aylin Caliskan. 2021. Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 122–133, Virtual Event USA. ACM.

Suchin Gururangan, Dallas Card, Sarah K. Dreier, Emily K. Gade, Leroy Z. Wang, Zeyu Wang, Luke

Zettlemoyer, and Noah A. Smith. 2022. Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection. *arXiv:2201.10474 [cs]*. ArXiv: 2201.10474.

Donna Haraway. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, 14(3):575–599.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Po Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack W. Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 375–385, Virtual Event Canada. ACM.

Jigsaw. 2017. Kaggle's Toxicity Comment Classification competition.

Xisen Jin, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, and Xiang Ren. 2021. On transferability of bias mitigation effects in language model fine-tuning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3770–3783, Online. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating Training Data Mitigates Privacy Risks in Language Models. *arXiv:2202.06539 [cs]*. ArXiv: 2202.06539.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR.

Erin M. Kerrison, Jennifer Cobbina, and Kimberly Bender. 2018. "Your Pants Won't Save You": Why Black Youth Challenge Race-Based Police Surveillance and the Demands of Black Respectability Politics. *Race and Justice*, 8(1):7–26.

Os Keyes, Chandler May, and Annabelle Carrell. 2021. You keep using that word: Ways of thinking about gender in computing research. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

William Labov. 1986. The social stratification of (r) in new york city department stores. In Harold B. Allen and Michael D. Linn, editors, *Dialect and Language Variation*, pages 304–329. Academic Press, Boston.

Mandy Lau. 2021. Artificial intelligence language models and the false fantasy of participatory language policies. *Working papers in Applied Linguistics and Linguistics at York*, 1:4–15.

Connor Leahy and Stella Biderman. 2021. The hard problem of aligning AI to human values. In *The State of AI Ethics Report (Volume 4)*. The Montreal AI Ethics Institute.

Josh Lepawsky. 2019. No insides on the outsides. *Discard Studies*, 0(0).

Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards Understanding and Mitigating Social Biases in Language Models.

In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6565–6576. PMLR.

Q. Vera Liao and Michael Muller. 2019. Enabling value sensitive ai systems through participatory design fictions.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *arXiv:2109.07958 [cs]*. ArXiv: 2109.07958.

Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Diptikalyan Saha. 2021. Generate your counterfactuals: Towards controlled counterfactual generation for text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13516–13524.

Liam Magee, Lida Ghahremanlou, Karen Soldatic, and Shanthi Robertson. 2021. Intersectional Bias in Causal Language Models. *arXiv:2107.07691 [cs]*. ArXiv: 2107.07691.

Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2021. Socially aware bias measurements for hindi language representations.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Angelina McMillan-Major, Zaid Alyafeai, Stella Biderman, Kimbo Chen, Francesco De Toni, Gérard Dupont, Hady Elsahar, Chris Emezue, Alham Fikri Aji, Suzana Ilić, Nurulaqilla Khamis, Colin Leong, Maraim Masoud, Aitor Soroa, Pedro Ortiz Suarez, Zeerak Talat, Daniel van Strien, and Yacine Jernite. 2022. Documenting Geographically and Contextually Diverse Data Sources: The BigScience Catalogue of Language Data and Resources. *arXiv:2201.10066 [cs]*. ArXiv: 2201.10066.

Milagros Miceli, Julian Posada, and Tianling Yang. 2022. Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP):1–14.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 220–229, Atlanta, GA, USA. Association for Computing Machinery.

Mike Monteiro and Vivianne Castillo. 2019. *Ruined by design: how designers destroyed the world, and what we can do to fix it*. Mule Design, Fresno.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Safiya Umoja Noble. 2018. *Algorithms of oppression: how search engines reinforce racism*. New York University Press, New York.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2021. Bbq: A hand-built bias benchmark for question answering.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the Everything in the Whole Wide World Benchmark. *arXiv:2111.15366 [cs]*. ArXiv: 2111.15366.

Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 429–435, Honolulu HI USA. ACM.

Micah Rajunov and Scott Duane. 2019. *Nonbinary: Memoirs of Gender and Identity*. Columbia University Press.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining Algorithmic Fairness in India and Beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 315–328, Virtual Event Canada. ACM.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask Prompted Training Enables Zero-Shot Task Generalization. *arXiv:2110.08207 [cs]*. ArXiv: 2110.08207.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2021. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. *arXiv:2111.07997 [cs]*. ArXiv: 2111.07997.

Daniel Schiff. 2020. Principles to Practices for Responsible AI: Closing the Gap. *2020 European Conference on AI Workshop on Advancing Towards the SDGs*.

Shanya Sharma, Manan Dey, and Koustuv Sinha. 2021. Evaluating gender bias in natural language inference.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal Biases in Language Generation: Progress and Challenges. In *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)*, pages 4275–4293, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3405–3410, Hong Kong, China. Association for Computational Linguistics.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. 2022. Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model. *arXiv:2201.11990 [cs]*. ArXiv: 2201.11990.

Dean Spade. 2015. *Normal Life: Administrative Violence, Critical Trans Politics, and the Limits of Law*. Duke University Press.

Karolina Stanczak and Isabelle Augenstein. 2021. A Survey on Gender Bias in Natural Language Processing. *arXiv:2112.14168 [cs]*. ArXiv: 2112.14168.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Arjun Subramonian. 2021. Allennlp: Fairness and bias mitigation.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Zeerak Talat, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. 2021. Disembodied Machine Learning: On the Illusion of Objectivity in NLP. ArXiv: 2101.11974.

Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. 2020. It's morphin' time! Combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2920–2935, Online. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew

Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications. *arXiv:2201.08239 [cs]*. ArXiv: 2201.08239.

Nanna Thylstrup and Zeerak Talat. 2020. Detecting 'Dirt' and 'Toxicity': Rethinking Content Moderation as Pollution Behaviour. *SSRN Electronic Journal*.

Nenad Tomasev, Kevin R. McKee, Jackie Kay, and Shakir Mohamed. 2021. *Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities*, page 254–265. Association for Computing Machinery, New York, NY, USA.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. *arXiv:1912.07076 [cs]*. ArXiv: 1912.07076.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5309–5318, Seoul, Korea (South). IEEE.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021a. Ethical and social risks of harm from Language Models. *arXiv:2112.04359 [cs]*. ArXiv: 2112.04359.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021b. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Sarah West, Meredith Whittaker, and Kate Crawford. 2019. Discriminating Systems: Gender, Race, and Power in AI. Technical report, AI Now Institute, New York.

Langdon Winner. 1980. Do artifacts have politics? *Daedalus*, pages 121–136.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining Gender Bias in Languages with Grammatical Gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5275–5283, Hong Kong, China. Association for Computational Linguistics.

Ran Zmigrod, Sabrina Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.