The Impossibility of Fair LLMs

Jacy Reese Anthis
University of Chicago

Kristian Lum
Google DeepMind

Michael Ekstrand
Drexel University

Avi Feller
University of California, Berkeley

Alexander D'Amour
Google Research

Chenhao Tan
University of Chicago

ABSTRACT

The need for fair AI is increasingly clear in the era of general-purpose systems such as ChatGPT, Gemini, and other large language models (LLMs). However, the increasing complexity of human-AI interaction and its social impacts have raised questions of how fairness standards could be applied. Here, we review the technical frameworks that machine learning researchers have used to evaluate fairness, such as group fairness and fair representations, and find that their application to LLMs faces inherent limitations. We show that each framework either does not logically extend to LLMs or presents a notion of fairness that is intractable for LLMs, primarily due to the multitudes of populations affected, sensitive attributes, and use cases. To address these challenges, we develop guidelines for the more realistic goal of achieving fairness in particular use cases: the criticality of context, the responsibility of LLM developers, and the need for stakeholder participation in an iterative process of design and evaluation. Moreover, it may eventually be possible and even necessary to use the general-purpose capabilities of AI systems to address fairness challenges as a form of scalable AI-assisted alignment.

KEYWORDS

large language models, natural language processing, deep learning, human-AI interaction, algorithmic fairness, algorithmic bias, fairness, bias, discrimination

1 INTRODUCTION

The rapid adoption of machine learning in the 2010s was accompanied by increasing concerns about negative societal impact, particularly in high-stakes domains. In response, there has been extensive development of technical frameworks to formalize ethical and social ideals—particularly the foundational notion of "fairness"—so that they can be evaluated and applied. Popular frameworks in machine learning and natural language processing (NLP) include group fairness [20] and fair representations [72]. In general, the frameworks we have today are oriented towards systems that are used in ways more-or-less self-evident from their design, typically with well-structured input and output, such as the canonical examples of predicting default in financial lending [39], predicting recidivism in criminal justice [4], and coreference resolution in natural language [74].

Recently, there has been a surge of interest in generative AI and general-purpose large language models (LLMs) that are pre-trained for next-word prediction and instruction-tuned to satisfy human preferences. LLMs are increasingly used for a multitude of tasks that span both traditional areas of concern for bias and fairness—such as evaluating resumes in hiring [6]—and areas less frequently, if at all, discussed in the extant fairness literature—such as drafting and editing emails [41], answering general knowledge queries [63], and code completion in software development [8].

In this paper, we consider whether and how the fairness frameworks can be applied to the LLM paradigm. We approach this topic mindful of both the hotly contested issues already present in the fairness literature [e.g., 17] and the challenges that other ascendant paradigms have presented for the ideal of fairness, such as information access systems [23]. For example, it is clear from the extant literature that achieving multiple fairness metrics simultaneously is generally intractable. Well-known impossibility results show that multiple group fairness metrics, such as those defined by rates of false positives and false negatives [16, 38] or demographic parity (Definition 3) and calibration (Definition 5) [38], cannot be simultaneously achieved in real-world environments. In this paper we develop a stronger claim that an LLM cannot achieve fairness even on a single non-trivial metric.

Before reviewing the recent work on LLM fairness, we ground our technosocial analysis in features of the LLM paradigm that seem essential for fairness evaluation. First, at the technical level, LLMs have exceptional flexibility. While the input and output have largely been restricted to natural language, it is increasingly clear that a wide range of content can be represented in LLM-suitable natural language. Moreover, LLMs, or more broadly the class of so-called "foundation models" [12], are increasingly multimodal, such as the ability of GPT-4 [54] to receive text, images, or combinations of the two as input. This flexibility is reflected in the lack of a self-evident use case or even a relatively narrow set of use cases—the existence of which has grounded technical fairness analysis in the past.

Second, at the social level, our analysis foregrounds the multitude of diverse stakeholders in LLM systems and their evolving relationships. As discussed in Section 3.2, there are people or organizations who create datasets, curate datasets, develop models, deploy and manage models, and build downstream user-facing applications—although there is often now a single organization that develops the model and plays many of the other roles. As with other information systems, there are always users—whether individuals or groups—who may have widely varying competencies [25]. Usually there are also subjects of the content produced by the system,

1st HEAL Workshop at CHI Conference on Human Factors in Computing Systems, May 12, 2024, Honolulu, HI, USA. Please cite the published version.

such as the people or groups that are described in an information request. There may also be researchers and critics from academia, governments, and nonprofit organizations analyzing LLM systems and their societal impacts and attempting to steer them in socially beneficial directions.

With these dynamics in mind, Section 2 discusses work to date on LLM fairness, which has focused on association-based metrics and practical challenges rather than the more nuanced metrics and inherent challenges we articulate in the present work. Section 3 shows that there is a fundamental, logical incompatibility between some extant frameworks and modern LLM systems. Section 4 shows that, in the other frameworks, the flexibility of LLMs across data, tasks, stakeholders, and populations renders a general guarantee, stamp, or certificate of a fair LLM intractable. Section 5 proposes moving forward in the important goals of fairness and harm reduction in LLMs by transmuting the inherent challenges into three general guidelines: the criticality of context, the responsibility of LLM developers, and the need for iterative and participatory design. We conclude with a practical discussion of what these guidelines imply for current LLM practices of curating training data, instruction tuning, prompt engineering, personalization, and using interpretability tools.

2 RECENT WORK ON LLM FAIRNESS

Interest in LLMs, particularly models based on the Transformer architecture introduced in 2017 [68], has greatly accelerated since 2020 with the popularity of OpenAI's GPT models [53] and more recently the proliferation of LLMs, such as Anthropic's Claude and Google's Gemini. A number of recent papers have discussed and evaluated bias, discrimination, and fairness in LLM-generated text.

2.1 Association-based fairness metrics

Two recent reviews of this nascent literature [30, 44] enumerate a variety of fairness metrics, each of which constitutes an association between a feature of the embedding space or model output (token probabilities or generated text) and a sensitive attribute. This includes text measures such as a disparity of sentiment and toxicity in Wikipedia sentence completion across the profession, gender, race, religion, or political ideology of the article subject [18], the occurrence of violent words after a phrase such as "Two muslims walked into a" [3], and the topics introduced when completing sentences from fiction novels [48]. Other approaches include creating datasets of LLM continuations of text that stereotypes, demeans, or otherwise harms in ways related to gender and sexuality [29]; evaluating conventional fairness metrics when an LLM is used for a conventional machine learning task, such as predicting outcomes based on a text-converted tabular dataset [45]; recommending music or movies to a user who specifies their sensitive attribute such as race or religion [73]; and testing whether the model provides the same "yes" or "no" answer when asked for advice by a user who specifies their gender [66]. In general, it would be very surprising if these models did not have disparate output given how they are trained, but these studies have provided useful, rigorous documentation.

However, a lack of such disparities, which have been the predominant target of study in LLM fairness studies to date, does not constitute fairness as conceptualized in machine learning or other fields such as philosophy Binns [7]. For example, in the framework of group fairness, which uses conditional equivalencies across sensitive attributes, the unconditional equivalence in model classification or prediction is known as demographic parity (see Definition 3). While demographic parity is an important metric for comprehensive fairness evaluation and enforcement, achieving it is rarely if ever viewed as achieving algorithmic fairness. In general, while the popular benchmarks such as WinoBias [74] and BBQ [56] that have been applied to LLM-generated text to date capture important information about model behavior in relation to sensitive attributes, there is little reason to think that strong model performance would imply fairness itself.

When existing work on LLMs has touched on richer notions of fairness, that has been undertaken in a highly constrained manner. For example, while Li et al. [44] briefly discussed counterfactual fairness (Definition 7), they only did so by summarizing two papers that merely perturb the LLM input (e.g., converting Standard American English to African American English [46]), which does not acknowledge or address the inherent challenges we present in Section 4.1 of how counterfactual fairness and other metrics fail to generalize across populations in which the data-generating process could vary significantly and counterfactuals would not merely vary in writing style or other features directly observable from the text.

2.2 Practical challenges

Existing work has motivated and articulated significant challenges in evaluating and enforcing fairness in LLMs. Both Gallegos et al. [30] and Li et al. [44] provide useful summaries, including the need to center marginalized communities through participatory design [9, 10] and to develop better proxy metrics, such as by bridging the divide between intrinsic and extrinsic bias metrics [32]. These are important challenges to be addressed, but even if each of them were, the inherent challenges that are the focus of the present work would remain.

The inherent challenges of LLM fairness have yet to be foregrounded in part because existing work has focused on relatively narrow use cases, often analyzing the LLM as a classifier or recommender system in conventional machine learning tasks through the use of in-context learning to steer the model towards the conventional output format (e.g., a binary data label or recommendation) [45, 66, 73]. It is true that, given the flexibility of LLMs as general text-to-text or multimodal models, they can be deployed—though not necessarily with strong performance—to any conventional task in which the input and output is a series of tokens or other suitable representation. However, LLMs are not primarily used as substitutes for conventional, narrow-purpose models. Examples of the rapidly growing set of use cases include coding (e.g., creation, autocompletion), communication (e.g., drafting emails, translation), gathering information (e.g., web search, proprietary search), recreation (e.g., bedtime stories, personalized travel plans), and simulation (e.g., data labeling, persona generation). Many of these tasks have not been rigorously considered in the extant fairness literature despite their increasing prevalence.

3 SOME FAIRNESS FRAMEWORKS CANNOT BE APPLIED TO LLMS

While the extant machine learning literature—prior to the popularization of LLMs—has shown that many fairness metrics are incompatible with each other [16, 38], here we develop the stronger claim that some frameworks cannot even be logically applied to a single metric in the case of an LLM.

3.1 Unawareness is impossible by design

Though often used as a strawman in the fairness literature, arguably the most common approach to algorithmic fairness has been fairness through unawareness (FTU).

Definition 1. (Fairness through unawareness). A model achieves fairness through unawareness (FTU) if the input to the model does not explicitly contain any sensitive attributes.

The concept of FTU emerged in the context of models built on structured data, typically in which data is organized into variables that are used for prediction or classification. For example, a financial lending model could use a person's age, gender, and credit score to make a prediction about loan repayment with FTU meaning "gender" is excised from the training data. Although it has been established that omitting the explicit sensitive attribute from a model at training time is insufficient to guarantee that unwanted correlations with that attribute do not exist in the model, legal, policy, or feasibility constraints often lead to this approach even though it is "widely considered naive" [71]. In one of the most widely known allegations of algorithmic discrimination, a group of heterosexual spouses who used the Apple Card to make purchases noticed after online discussion that each woman was extended a much lower credit limit than her husband. The company managing the Apple Card, Goldman Sachs, defended itself by saying, "In all cases, we have not and will not make decisions based on factors like gender" [67]. The primary critique of this approach is that the sensitive attribute is, often strongly, related to other attributes included in the training data, so models effectively recover the excluded sensitive attribute and this curtails any potential fairness benefits.

By design, LLMs are trained on unstructured modalities such as natural language, a context in which FTU is impossible because of the pervasiveness of sensitive attributes. Indeed, LLMs are readily able to infer personal characteristics such as age, location, and gender of an author from their written text [64]. Efforts to remove sensitive attributes from text may produce incoherence or distortion. For example, in the sentence, "Alice grew up in Portugal, so Alice had an easy time on the trip to South America," simply removing Alice's national origin of "Portugal" would result in an ungrammatical sentence. Other approaches for removing national origin would still result in distortion. Substituting the neutral phrase "a country" would remove important narrative information, such as the author conveying to the reader that Alice visited Brazil, the only South American country in which Portuguese is an official language.

Consider how the relative social status of characters in a narrative can be conveyed through pronoun usage in quoted material, such as the more frequent use of first-person pronouns being more common in groups of lower social status [35]. Moreover, in languages with gendered nouns (e.g., Spanish, German), enforcing gender fairness may require introducing entirely new vocabulary, and if nationality, native language, religion, beliefs, or other attributes of cultural background are considered sensitive, then the corresponding languages, dialects, and subdialects would also be impossible to extirpate. Even with attributes that could be removed without distortion in certain cases, it is infeasible to enforce fairness with respect to all relevant sensitive attributes across a large corpus while retaining sufficient information for model performance. There may also be direct ethical issues with the modification of text, such as individual authors not consenting to the modification of text they own.

As with the other frameworks, FTU is additionally hindered by the current lack of model transparency. FTU would require that the LLM be documentably unaware of the sensitive information, which requires a level of documentation of training data that is unavailable from any state-of-the-art LLM today—at least to third-party researchers, auditors, and developers. Finally, while conventional FTU explicitly leaves out the sensitive attribute, some approaches use the sensitive attribute information to ensure that the model is not even implicitly aware of the sensitive attribute through proxies, such as zip code as a proxy for race and income [47, 57]. The current lack of LLM documentation further prevents researchers and socially aware application developers from studying model awareness of sensitive attributes in the first place.

3.2 LLMs can render producer-side fairness criteria obsolete

In the literature on fairness in recommender and information retrieval systems, the presence of multiple stakeholders has motivated a framework of multi-sided fairness. In this section, we consider each type of stakeholder in turn, highlighting the existing difficulties of fairness for each type and showing that each challenge compounds in the case of LLMs—particularly in the case of content producers because LLMs can extract content and present it to users with little or no compensation to its producer. Stakeholders are typically divided into the consumers, producers, and subjects of content [2, 13, 23, 62].

Definition 2. (Multi-sided fairness). A system achieves multi-sided fairness if the model is fair with respect to each group of its stakeholders, such as the consumers of its content (C-fairness), the producers of its content (P-fairness), the consumers and producers of its content considered together (CP-fairness), and the subjects of the items that are being provided (S-fairness).

For users or consumers (i.e., the people or groups who receive the recommendations), there are many possible fairness targets, such as that each consumer or consumer group should receive comparably high-quality recommendations [22, 24, 26, 51, 69]. This target can more or less straightforwardly be applied to LLMs with the aforementioned challenges of a large diversity of use cases and user populations.

For subjects of LLMs, it may be difficult to define, detect, or enforce appropriate fairness metrics. For example, early in the literature on fairness in information access systems, it was noted that when searching for images of "CEO," Google returned a set of images largely depicting men and, lower in the recommendation list, an image of the popular toy "CEO Barbie." However, as in the FTU decision of which sensitive attributes a system should be unaware of and in what way, the challenges of deciding subject representation in system output are compounded with LLMs. These decisions typically rely on utility estimates, and that tends to be a significant challenge with more general-purpose systems based on unstructured data. For example, there is an open question of whether the target distribution should be equal representation of men, women, and other genders or a distribution that is weighted towards the gender distribution of CEOs in the consumer's home location [28, 37, 59]. To achieve LLM fairness, this sort of open question would need to be resolved for each of the many tasks done by the LLM.

For producers (i.e., the people or organizations whose content is recommended), also known as providers, the fairness target is often the equitable distribution of exposure, either in terms of relevance-free metrics that do not consider the relevance of the content to the user—only that there is an equitable distribution—or relevance-based fairness metrics that target an equitable exposure conditional on relevance. In either case, fairness to producers is a matter of how the exposure of those providing content to the system is allocated to consumers. In the use case of LLMs that perform information retrieval and information management tasks, this framework can at times transfer directly. For example, if someone searches for "coffee shops in San Francisco" in an LLM chat or search interface—as is being incorporated into the ubiquitous modern search engine, Google—producer fairness could be defined in terms of equitable exposure to the different brick-and-mortar coffee shops in San Francisco. Even if the LLM system does not direct users to particular websites, many users will presumably end up visiting the cafes, which provides utility—fairly or unfairly—to the producers. However, if users are searching for information in the LLM system, such as asking, "How are coffee beans roasted?" then LLMs can entirely circumvent the producers and upend the conventional notion of producer-side fairness. If the LLM system extracts information from websites without directing users to the original source content, then it may be that none of the producers receive any exposure or other benefits in the first place. One way to make sense of this would be to consider the LLM system itself—or the entity that developed, owns, and manages it—as another type of stakeholder, one that takes all utility from the producers and renders the conventional producer-side fairness criteria obsolete.

4 LLMS ARE TOO FLEXIBLE TO BE GENERALLY FAIR

Much of the excitement surrounding LLMs is based on their general-purpose flexibility across wide ranges of inputs, tasks, outputs, and contexts. To some extent, they resemble a human agent, including the ability to chain together these tasks into complex sequences, and these areas of flexibility make many conventional fairness metrics intractable.

4.1 Group fairness does not generalize across populations

Group fairness metrics require independence between model classification and sensitive attributes, often conditional on relevant information such as the ground-truth labels that the model aims to predict (e.g., job performance for a model that assists in hiring decisions). Three common metrics are:

Definition 3. (Demographic parity). A model achieves demographic parity if its predictions are statistically independent of sensitive attributes.

Definition 4. (Equalized odds). A model achieves equalized odds if its predictions are statistically independent of sensitive attributes conditional on the true labels being predicted.

Definition 5. (Calibration). A model achieves calibration if the true labels being predicted are statistically independent of sensitive attributes conditional on the model's predictions.

In binary classification, these metrics are achieved when equalities hold between ratios in the confusion matrix: equal ratios of predicted outcomes (demographic parity), equal true positive rates and false positive rates (equalized odds), and equal precision (calibration). Recent work includes extensions of these notions, such as prioritizing the worst-off group by minimizing the maximum group error rate [19]. Conventionally, group fairness requires knowing the sensitive attributes to enforce the equalities, though recent work has considered approaches for when the sensitive attributes are unavailable [36, 42, 76]. There are many methods for enforcing group fairness metrics, such as the preprocessing of datasets proposed by Feldman et al. [27] to guarantee bounds on demographic parity and the more recent method proposed by Johndrow and Lum [34] that can be applied to a wider variety of datasets.

LLMs present a challenge for group fairness metrics in part because LLMs tend to be deployed across a wide range of data distributions. Lechner et al. [43] showed that it is impossible for a non-trivial model to perform fairly across all different data distributions, such as regions or demographic groups, to which it might be applied. In current discussions of algorithmic fairness (e.g., recidivism), fairness is typically targeted at a local jurisdiction, which ensures that the model is performing fairly for that location's particular demographic mix and holistic characteristics but typically cannot also achieve fairness in substantially different locations. The purpose and use of LLMs makes it infeasible to restrict them to this sort of targeted population.

In general, it is not clear what an appropriate base population would be on which to detect and achieve group fairness for an LLM. For example, one could "bootstrap" a predictive model for recidivism prediction from an LLM simply by instructing it to make a prediction about an individual based on a fixed set of that individual's characteristics with in-context learning as Li and Zhang [45] did in predicting the label of a text-converted tabular dataset. However, the data on which that LLM had been trained does not admit an identifiable base population because a corpus of text is not a structured database comprising people and their characteristics. An LLM may be trained in part on such databases, but the output of the model for such predictions will also be based on the wide scope of unstructured natural language or other modalities of data on which the model is trained.

Generalization across populations is also a concern for fairness frameworks other than group fairness because of the wide range of data, use cases, and social contexts at play in LLMs [60]. Here, we consider two examples: individual fairness [20] and counterfactual fairness, which is the most common causal notion of fairness [40].

Definition 6. (Individual fairness). A model achieves individual fairness if similar individuals are treated similarly. Formally, this requires that the distribution of model output is Lipschitz continuous with respect to the distribution of model input.

Definition 7. (Counterfactual fairness). A model achieves counterfactual fairness if the model would produce the same output for an individual if they had a different level of the sensitive attribute.

In terms of individual fairness, it is not clear what similarity metrics could be reasonably applied across the multitude of contexts or, if multiple metrics were applied, how these could be judiciously selected and guaranteed in each possible context. In terms of causal fairness, including counterfactual fairness, it is often difficult to identify the causal structure of the data-generating process in even a single NLP task, and it would be an immense challenge for a single model to account for all of the many different contextual factors that determine counterfactuals or other causally distinct outcomes across the varying populations.

4.2 Sensitive attributes proliferate in a general-use setting

The preceding section considered the challenges of imposing fairness across different data distributions. When considering different sensitive attributes, given the issues discussed in Section 3.1, it may not be tractable to exclude sensitive attributes from the training data, and each of the different distributions and different tasks can require fairness metrics to be enforced for a different set of sensitive attributes. This is a challenge for the group fairness metrics already defined, but the issue is particularly salient for the popular ideal of fair representations within a machine learning model or fair representations produced by one model and used by another [72].

Definition 8. (Fair representation). A representation is fair if it does not contain information that can identify the sensitive attributes of the individuals being represented.

In the fair representations framework, a system first maps the dataset of individuals being represented to a probability distribution in a novel representation space, such that the system preserves as much information as possible about the individual while removing all information about the individual's sensitive attribute. The most well-known example of this approach is Bolukbasi et al. [11], which rigorously documented gender bias in Google News word embeddings, namely an association between occupations and a gender vector (e.g., $\vec{he} - \vec{she}$), such that computer programmer was coded as highly male while homemaker was coded as highly female. Indeed, this is where much of the NLP fairness literature has focused, documenting similar biases across different word embedding models Sesari et al. [see 61, for a review].

Researchers have developed a number of debiasing approaches focused on the sensitive attribute dimension, such as zeroing the projection of each word vector (e.g., each occupation) onto the dimension itself [11] or training the model to align the sensitive attribute dimension with the last coordinate of the embedding space, so that it can be easily removed or ignored [75]. However, Gonen and Goldberg [33] show that such approaches "are mostly hiding the bias rather than removing it" because, after removal, word pairs tend to maintain their similarity, which still reflects associations with sensitive attributes—what Bolukbasi et al. [11] call "indirect bias."

Achieving fairness in one LLM context may be contingent on the removal of information or alteration of the statistical relationships between the context-specific sensitive attribute and other features of the data. For example, one may wish to exclude gender information from financial lending decisions, but gender information may be necessary for other tasks, such as drafting or editing an email about a real-world situation that has important gender dynamics that the sender hopes to communicate to the receiver. Moreover, variables highly correlated with gender, such as biological sex and pregnancy status, may be essential criteria for medical decision-making. In general, attempts at debiasing for one context may remove or distort important information for another context.

The naive approach of debiasing the model with respect to the union of all potential sensitive attributes—even if it were empirically feasible—would likely be too heavy-handed, leaving the model with little information to be useful for any task. To effectively create a fair LLM for every task and context, one would need to act upon the parameters of the model with surgical precision to alter the relationship between variables only when the model is instantiated for a specific task and context. This is infeasible with current LLM methods, such as fine-tuning, and currently we do not have robust techniques to debias even a single problematic relationship without incidentally obfuscating it or problematizing other relationships. This game of fairness whack-a-mole seems indefinitely intractable. Likewise, even if we could reduce the union of all potential sensitive attributes to a manageable level, such as identifying a small set of the most important to adjust for in each task, that would still require yet-infeasible fine-grained adjustments to avoid counterproductive side effects.

4.3 Fairness does not compose, but fairness-directed composition may help

Whether a model's behavior on a given task is fair or desirable largely depends on how the model's output will be used. In modern AI systems including LLMs, the output of one model is often used as the input to another model, but this produces an additional challenge because fairness does not compose: a fairness guarantee for each of two models is not a fairness guarantee for a system composed of the two models—a point made most explicitly by Dwork and Ilvento [21].

However, it is this inherent challenge of LLM fairness that most directly suggests a productive direction for the LLM fairness literature. Namely, it may be possible to use the aforementioned flexibility of general-purpose LLMs to effectively create fair context-specific model compositions, enforcing fairness ideals in seemingly intractable contexts. This is due to, first, the LLMs' ability to account for many patterns in data not immediately observable by human model designers—which is much of the reason for excitement about LLMs in recent years—and, second, the instruction tuning that allows them to obey natural language input. Eventually, they may be able to obey a general command to enforce context-specific fairness. Many advances in LLM capabilities can be conceptualized as encouraging the model to improve its own output. For example, chain-of-thought prompting [70] encourages the model to first produce text that takes an incremental reasoning step towards its target, which can increase performance by allowing the later token generations to build on the logical reasoning text that the model has already generated, which has then become part of its input.

In terms of fairness and other ethical issues, one can view many approaches to instruction tuning as a composition of an ethics-driven model with the primary LLM. The most popular approaches to alignment and safety, currently Reinforcement Learning from Human Feedback [RLHF; 55] and Direct Preference Optimization [DPO; 58], compel the model towards human-provided preference data, and some other approaches, such as Constitutional AI [5] and SELF-ALIGN [65], steer the model towards LLM-generated proxies of human preferences.

While AI-assisted fairness is an interesting possibility, it could easily make the situation worse if attempted before models have the capability to do this safely. The fairness-enforcing model could double down on its own blindspots, particularly those that are not yet sufficiently well-understood or appreciated by the human developers such that they can be guarded against. Recent approaches focus on model "self-correction." While there is skepticism that models can currently do this well, Ganguli et al. [31] show impressive results on bias and discrimination benchmarks "simply by instructing models to avoid harmful outputs." As LLM capabilities rapidly increase in the coming years, scalable approaches that allow us to leverage those capabilities for fairness may be necessary despite the risks.

5 MOVING FORWARD

Over the past decade, the machine learning community has developed several compelling technical frameworks for assessing the impact of machine learning methods deployed in high-stakes domains. We have shown that current frameworks are insufficient for the critical task of addressing fairness issues in the case of LLMs. It is not feasible to certify or guarantee that an LLM is generally "fair," in part because of the inapplicability of some frameworks to LLM tasks, including unstructured natural language data, and in part because of the intractability of enforcing fairness on flexible, general-purpose foundational models, such as LLMs. Here we develop three guidelines to move forward on LLM fairness and discuss tentative implications for specific LLM practices.

5.1 Guidelines

For researchers evaluating the societal impacts of LLMs, context is critical. A key strength of LLMs is that the same model can be fine-tuned for many different applications in many different contexts. Making meaningful statements about LLMs behaving fairly—even if we can't say that they are generally fair—will require articulating connections to real use cases and corresponding harms. Fairness evaluations must reflect the diversity of these contexts, expanding beyond the nascent LLM fairness literature that has primarily focused on decontextualized, hypothetical, and ungrounded tests of associations between sensitive attributes and model output. With the challenges in translating and composing fairness across models and domains, it is unlikely that any "trick tests," such as coreference resolution of gendered pronouns, will provide satisfactory evidence for or against LLM fairness [50]. There has been a dearth of proper contextualization for years in the fairness literature [10], and the rise of LLMs makes this even more concerning.

LLM developers are responsible for safe use and harm mitigation. Fairness is necessarily a feature of end-to-end pipelines from model design and training to model deployment and long-term effects. While users, regulators, researchers, and auditors have historically been well-positioned to collect and evaluate data on the later stages of this pipeline, there are substantial challenges in understanding and managing the earlier stages that are necessary for comprehensive solutions. LLM developers have a responsibility to empower stakeholders to assess fairness of LLM-based applications in these varied contexts. Most immediately, for researchers and other third parties to move beyond ungrounded prompts and contexts, companies that deploy LLMs, such as OpenAI and Google, must share information on actual usage and how the systems respond to real prompts from real users [14, 50]. LLM developers also have a responsibility to support these efforts through technical training, tools to facilitate evaluation of specific tasks, and other resources.

Managing the societal impact of LLMs will require iterative and participatory design and evaluation. Given the many different contexts in which LLMs are used, which have conflicting desiderata for defining fairness, LLM developers must work closely with third-party researchers, policymakers, end users, and other affected stakeholders in a participatory design process [52] that audits algorithms in the contexts in which they are used and mitigates harmful effects. Given the novel fairness challenges of LLMs and the amplification of existing challenges, this process must also be iterative with frequent trials and assessments of harm. While we have provided reason for skepticism of current approaches, there is still ample room for fair and responsible AI development as evaluated in particular use cases.

5.2 Implications

The recent surge of interest in LLMs has led to the development of many common LLM practices. Here we consolidate the concrete implications of the inherent fairness challenges for such practices. However, the usage of LLMs has been rapidly evolving as people experiment with new applications of increasingly powerful models, and while we attempt to focus on the general practices that will likely continue in some form, future work may need to revisit the aforementioned guidelines and update the particular implications accordingly.

Training data. Modern generative AI systems are trained with unprecedented amounts of natural language and multimodal data. One of the most frequently discussed issues raised by the extensive data scraping used to collect such large datasets is copyright and intellectual property law [1]. In addition to existing concerns in machine learning, such as biases within datasets, LLMs manifest unique challenges as the lack of developer transparency prevents user and third-party evaluation and as the value created by producers of training data, as with other content, can be extracted without credit by the LLM system. Fairness in LLMs will require increased documentation of LLM training data and consideration of how data representing different populations in different contexts from different perspectives (e.g., a person writing about their own experiences or an outside journalist documenting that person's life) is included and weighted. This includes ways that "quality" text filters may disproportionately exclude certain people [49].

Instruction tuning. We discussed how ethics-driven models can be used to steer LLMs in beneficial directions, such as the reward model in RLHF. However, fairness challenges can manifest in the process of instruction tuning itself, including feedback gathered from users when the LLM is used in practice. For example, it is not feasible to enforce a fairness standard across the diverse population whose preferences are considered by the model or even the preferences of the same people over time [15]. Just as a condition such as group fairness could not be enforced across the model output, it could not be enforced across the effects of user feedback on the LLM. It will be important to consider what distributions of feedback are most desirable given the many stakeholders who could give feedback and the many use cases on which that feedback could be targeted. Each approach to collecting and incorporating human input corresponds to particular value judgments.

Prompt engineering. The interactive feedback loops of LLM prompt rewriting and engineering constitute another threat vector in which bias can manifest even if the final generated text achieves fairness by one of the aforementioned metrics. For example, users who speak low-resource languages may face compounded challenges from a more limited ability to iterate on their prompting strategies in addition to the lower model performance. Like instruction tuning, prompt engineering can also be a tool to achieve better fairness outcomes through iteration and modification of LLM use.

Personalization. In practice, LLM-hosting platforms increasingly customize the system for individual users, such as by prepending the user's past chat history as context for the current output, and allow users to customize the system, such as writing custom instructions that are similarly prepended. As with instruction tuning and prompt engineering, this is a vector for both fairness issues (e.g., past harmful LLM output is now perpetuated by inclusion in the chat history) and opportunities (e.g., custom instructions that compel the LLM to treat users more fairly).

Interpretability tools. Various tools have been developed to interpret the behavior and internal mechanisms of an LLM. Bias can manifest through the use of interpretability tools, such as the quality of interpretation provided to different users in different contexts. For example, LLMs are typically trained to refuse to discuss particularly sensitive topics, and if we successfully build tools that can at times elicit faithful explanations of past behavior from a model, models may therefore refuse to provide accurate explanations of their own behavior when certain contexts and populations are involved. On the other hand, if interpretability tools succeed in providing relevant information, this can be used to make fairer models, such as if the internal mechanisms involve certain associations between sensitive attributes and important variables and this can be modified or incorporated into fair practices. Still, interpretability tools should be treated with substantial caution, at least for the immediate future, because the appearance of relevant information may be misleading. For example, RLHF may encourage the model to incorrectly report that it has implemented fairness practices because that is a response users express a revealed preference for during training because the users are not aware that it is incorrect.

In summary, even though achieving fairness is difficult, it is essential for responsible LLM development and deployment. With an iterative approach grounded in the nature of LLMs, real-world use cases, and a cautious use of AI tools themselves for fairness evaluation and enforcement, we believe that substantial progress could be made in a relatively short period of time.

REFERENCES

[1] Ryan Abbott and Elizabeth Rothman. 2023. Disrupting Creativity: Copyright Law in the Age of Generative Artificial Intelligence. *Florida Law Review* 75, 6 (30 11 2023), 1141–1201.

[2] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. 2020. Multistakeholder Recommendation: Survey and Research Directions. *User Modeling and User-Adapted Interaction* 30, 1 (March 2020), 127–158. https://doi.org/10.1007/s11257-019-09256-1

[3] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Large Language Models Associate Muslims with Violence. *Nature Machine Intelligence* 3, 6 (June 2021), 461–463. https://doi.org/10.1038/s42256-021-00359-2

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *ProPublica* (2016).

[5] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073 [cs]

[6] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review* 94, 4 (Sept. 2004), 991–1013. https://doi.org/10.1257/0002828042002561

[7] Reuben Binns. 2021. Fairness in Machine Learning: Lessons from Political Philosophy. *arXiv:1712.03586 [cs]* (March 2021). arXiv:1712.03586 [cs]

[8] Christian Bird, Denae Ford, Thomas Zimmermann, Nicole Forsgren, Eirini Kalliamvakou, Travis Lowdermilk, and Idan Gazit. 2022. Taking Flight with Copilot: Early Insights and Opportunities of AI-powered Pair-Programming Tools. *Queue* 20, 6 (Dec. 2022), 35–57. https://doi.org/10.1145/3582083

[9] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to

the People? Opportunities and Challenges for Participatory AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, Arlington VA USA, 1–8. https://doi.org/10.1145/3551624.3555290

[10] 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*. Association for Computational Linguistics, Online, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

[11] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings. *arXiv:1607.06520 [cs, stat]* (July 2016). arXiv:1607.06520 [cs, stat]

[12] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258 [cs]

[13] Robin Burke. 2017. Multisided Fairness for Recommendation. In *Proceedings of the Workshop on Fairness, Accountability and Transparency in Machine Learning (FATML)*. Halifax, NS, Canada. arXiv:1707.00093 [cs]

[14] Aylin Caliskan and Kristian Lum. 2024. Effective AI Regulation Requires Understanding General-Purpose AI. https://www.brookings.edu/articles/effective-ai-regulation-requires-understanding-general-purpose-ai/.

[15] Micah Carroll, Davis Foote, Anand Siththaranjan, Stuart Russell, and Anca Dragan. 2024. AI Alignment with Changing and Influenceable Reward Functions. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[16] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (June 2017), 153–163. https://doi.org/10.1089/big.2016.0047

[17] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Halifax NS Canada, 797–806. https://doi.org/10.1145/3097983.3098095

[18] 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*. 862–872. https://doi.org/10.1145/3442188.3445924 arXiv:2101.11718 [cs]

[19] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. 2021. Minimax Group Fairness: Algorithms and Experiments. arXiv:2011.03108 [cs]

[20] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. 2011. Fairness Through Awareness. *arXiv:1104.3913 [cs]* (Nov. 2011). arXiv:1104.3913 [cs]

[21] Cynthia Dwork and Christina Ilvento. 2019. Fairness Under Composition. *Leibniz International Proceedings in Informatics* (2019), 20 pages, 627743 bytes. https://doi.org/10.4230/LIPICS.ITCS.2019.33

[22] Michael D. Ekstrand, Lex Beattie, Maria Soledad Pera, and Henriette Cramer. 2024. Not Just Algorithms: Strategically Addressing Consumer Impacts in Information Retrieval. In *Proceedings of the 46th European Conference on Information Retrieval*.

[23] Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. 2022. Fairness in Information Access Systems. *Foundations and Trends® in Information Retrieval* 16, 1-2 (2022), 1–177. https://doi.org/10.1561/1500000079

[24] Michael D. Ekstrand and Maria Soledad Pera. 2022. Matching Consumer Fairness Objectives & Strategies for RecSys. arXiv:2209.02662 [cs]

[25] Michael D. Ekstrand, Maria Soledad Pera, and Katherine Landau Wright. 2023. Seeking Information with a More Knowledgeable Other. *Interactions* 30, 1 (Jan. 2023), 70–73. https://doi.org/10.1145/3573364

[26] Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All the Cool Kids, How Do They Fit in?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 172–186.

[27] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Sydney NSW Australia, 259–268. https://doi.org/10.1145/2783258.2783311

[28] Yunhe Feng and Chirag Shah. 2022. Has CEO Gender Bias Really Been Fixed? Adversarial Attacking and Improving Gender Fairness in Image Search. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 11 (June 2022), 11882–11890. https://doi.org/10.1609/aaai.v36i11.21445

[29] Eve Fleisig, Aubrie Amstutz, Chad Atalla, Su Lin Blodgett, Hal Daumé Iii, Alexandra Olteanu, Emily Sheng, Dan Vann, and Hanna Wallach. 2023. FairPrism: Evaluating Fairness-Related Harms in Text Generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 6231–6251. https://doi.org/10.18653/v1/2023.acl-long.343

[30] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2023. Bias and Fairness in Large Language Models: A Survey. *arXiv preprint arXiv:2309.00770* (2023). arXiv:2309.00770

[31] Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I. Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, Dawn Drain, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jackson Kernion, Jamie Kerr, Jared Mueller, Joshua Landau, Kamal Ndousse, Karina Nguyen, Liane Lovitt, Michael Sellitto, Nelson Elhage, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robert Lasenby, Robin Larson, Sam Ringer, Sandipan Kundu, Saurav Kadavath, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, Christopher Olah, Jack Clark, Samuel R. Bowman, and Jared Kaplan. 2023. The Capacity for Moral Self-Correction in Large Language Models. arXiv:2302.07459 [cs]

[32] Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2020. Intrinsic Bias Metrics Do Not Correlate with Application Bias. *arXiv preprint arXiv:2012.15859* (2020). arXiv:2012.15859

[33] 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*. Association for Computational Linguistics, Minneapolis, Minnesota, 609–614. https://doi.org/10.18653/v1/N19-1061

[34] James E. Johndrow and Kristian Lum. 2019. An Algorithm for Removing Sensitive Information: Application to Race-Independent Recidivism Prediction. *The Annals of Applied Statistics* 13, 1 (2019), pp. 189–220. jstor:26666180

[35] Ewa Kacewicz, James W. Pennebaker, Matthew Davis, Moongee Jeon, and Arthur C. Graesser. 2014. Pronoun Use Reflects Standings in Social Hierarchies. *Journal of Language and Social Psychology* 33, 2 (March 2014), 125–143. https://doi.org/10.1177/0261927X13502654

[36] Nathan Kallus, Xiaojie Mao, and Angela Zhou. 2021. Assessing Algorithmic Fairness with Unobserved Protected Class Using Data Combination. *Management Science* (2021). https://doi.org/10.1287/mnsc.2020.3850

[37] Chen Karako and Putra Manggala. 2018. Using Image Fairness Representations in Diversity-Based Re-ranking for Recommendations. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*. ACM, Singapore Singapore, 23–28. https://doi.org/10.1145/3213586.3226206

[38] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. *Proceedings of Innovations in Theoretical Computer Science (ITCS), 2017* (Sept. 2016). https://doi.org/10.48550/arXiv.1609.05807

[39] I. Elizabeth Kumar, Keegan E. Hines, and John P. Dickerson. 2022. Equalizing Credit Opportunity in Algorithms: Aligning Algorithmic Fairness Research with U.S. Fair Lending Regulation. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Oxford United Kingdom, 357–368. https://doi.org/10.1145/3514094.3534154

[40] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc.

[41] Philippe Laban, Jesse Vig, Marti A. Hearst, Caiming Xiong, and Chien-Sheng Wu. 2023. Beyond the Chat: Executable and Verifiable Text-Editing with LLMs. arXiv:2309.15337 [cs]

[42] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H. Chi. 2020. Fairness without Demographics through Adversarially Reweighted Learning. arXiv:2006.13114 [cs, stat]

[43] Tosca Lechner, Shai Ben-David, Sushant Agarwal, and Nivasini Ananthakrishnan. 2021. Impossibility Results for Fair Representations. arXiv:2107.03483 [cs, stat]

[44] Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2024. A Survey on Fairness in Large Language Models. *Procedia Computer Science* 00 (2024), 1–28. arXiv:2308.10149 [cs]

[45] Yunqi Li and Yongfeng Zhang. 2023. Fairness of ChatGPT. arXiv:2305.18569 [cs]

[46] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research* (2023).

[47] Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. 2018. Does Mitigating MLs Impact Disparity Require Treatment Disparity?. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc.

[48] Li Lucy and David Bamman. 2021. Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding*, Nader Akoury, Faeze Brahman, Snigdha Chaturvedi, Elizabeth Clark, Mohit Iyyer, and Lara J. Martin (Eds.). Association for Computational Linguistics, Virtual, 48–55. https://doi.org/10.18653/v1/2021.nuse-1.5

[49] Li Lucy, Suchin Gururangan, Luca Soldaini, Emma Strubell, David Bamman, Lauren Klein, and Jesse Dodge. 2024. AboutMe: Using Self-Descriptions in Webpages to Document the Effects of English Pretraining Data Filters. *arXiv* (16 Jan 2024). http://arxiv.org/abs/2401.06408

[50] Kristian Lum, Jacy Reese Anthis, Chirag Nagpal, and Alexander D'Amour. 2024. Bias in Language Models: Beyond Trick Tests and Toward RUTEd Evaluation. arXiv:2402.12649 [cs, stat]

[51] Rishabh Mehrotra, Ashton Anderson, Fernando Diaz, Amit Sharma, Hanna Wallach, and Emine Yilmaz. 2017. Auditing Search Engines for Differential Satisfaction Across Demographics. In *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*. ACM Press, Perth, Australia, 626–633. https://doi.org/10.1145/3041021.3054197

[52] Michael J. Muller and Sarah Kuhn. 1993. Participatory Design. *Commun. ACM* 36, 6 (June 1993), 24–28. https://doi.org/10.1145/153571.255960

[53] OpenAI. 2022. Introducing ChatGPT.

[54] OpenAI. 2023. GPT-4V(Ision) System Card. https://cdn.openai.com/papers/GPTV_System_Card.pdf.

[55] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training Language Models to Follow Instructions with Human Feedback. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 27730–27744.

[56] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. BBQ: A Hand-Built Bias Benchmark for Question Answering. arXiv:2110.08193 [cs.CL]

[57] Devin G Pope and Justin R Sydnor. 2011. Implementing Anti-Discrimination Policies in Statistical Profiling Models. *American Economic Journal: Economic Policy* 3, 3 (Aug. 2011), 206–231. https://doi.org/10.1257/pol.3.3.206

[58] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model Is Secretly a Reward Model. arXiv:2305.18290 [cs]

[59] Amifa Raj and Michael D. Ekstrand. 2022. Fire Dragon and Unicorn Princess; Gender Stereotypes and Children's Products in Search Engine Responses. arXiv:2206.13747 [cs]

[60] Maribeth Rauh, John Mellor, Jonathan Uesato, Po-Sen Huang, Johannes Welbl, Laura Weidinger, Sumanth Dathathri, Amelia Glaese, Geoffrey Irving, Iason Gabriel, William Isaac, and Lisa Anne Hendricks. 2022. Characteristics of Harmful Text: Towards Rigorous Benchmarking of Language Models. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 24720–24739.

[61] Emeralda Sesari, Max Hort, and Federica Sarro. 2022. An Empirical Study on the Fairness of Pre-trained Word Embeddings. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Association for Computational Linguistics, Seattle, Washington, 129–144. https://doi.org/10.18653/v1/2022.gebnlp-1.15

[62] Nasim Sonboli, Robin Burke, Michael Ekstrand, and Rishabh Mehrotra. 2022. The Multisided Complexity of Fairness in Recommender Systems. *AI Magazine* 43, 2 (June 2022), 164–176. https://doi.org/10.1002/aaai.12054

[63] Sofia Eleni Spatharioti, David M. Rothschild, Daniel G. Goldstein, and Jake M. Hofman. 2023. Comparing Traditional and LLM-based Search for Consumer Choice: A Randomized Experiment. arXiv:2307.03744 [cs]

[64] Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. 2024. Beyond Memorization: Violating Privacy via Inference with Large Language Models. In *Proceedings of the International Conference on Learning Representations*. ICLR.

[65] Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-Driven Self-Alignment of Language Models from Scratch with Minimal Human Supervision. arXiv:2305.03047 [cs]

[66] Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. 2023. Evaluating and Mitigating Discrimination in Language Model Decisions. arXiv:2312.03689 [cs]

[67] Taylor Telford. 2019. Apple Card Algorithm Sparks Gender Bias Allegations against Goldman Sachs. *Washington Post* (Nov. 2019).

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. arXiv:1706.03762 [cs]

[69] Lequn Wang and Thorsten Joachims. 2021. User Fairness, Item Fairness, and Diversity for Rankings in Two-Sided Markets. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. ACM, Virtual Event Canada, 23–41. https://doi.org/10.1145/3471158.3472260

[70] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 24824–24837.

[71] Alice Xiang. 2021. Reconciling Legal and Technical Approaches to Algorithmic Bias.

[72] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. *Proceedings of Machine Learning Research* 28, 3 (2013), 325–333.

[73] Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. Is ChatGPT Fair for Recommendation? Evaluating Fairness in Large Language Model Recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 993–999. https://doi.org/10.1145/3604915.3608860 arXiv:2305.07609 [cs]

[74] 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)*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 15–20. https://doi.org/10.18653/v1/N18-2003

[75] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning Gender-Neutral Word Embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 4847–4853. https://doi.org/10.18653/v1/D18-1521

[76] Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. 2022. Towards Fair Classifiers Without Sensitive Attributes: Exploring Biases in Related Features. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 1433–1442. https://doi.org/10.1145/3488560.3498493 arXiv:2104.14537 [cs]