Dubious Debiasing: Inherent Challenges in Achieving Fairness in Large Language Models

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

Researchers have evaluated fairness in machine learning with a variety of technical frameworks, such as group fairness and fair representations. With the increasingly complex ways in which generative AI interfaces with human society, it is not clear how these frameworks can be extended to general-purpose systems, such as ChatGPT, Gemini, and other large language models (LLMs). Despite the critical importance of evaluating LLM fairness, we articulate inherent challenges. In some cases, extant frameworks cannot be applied to human-LLM interaction, and in others, the notion of a fair LLM is intractable due to the exceptional flexibility of LLMs in performing many different types of tasks with effects on a multitude of diverse stakeholders, including widely varying user populations. We conclude with motivating principles for fairness in LLM systems that foreground the criticality of context, the responsibility of LLM developers, and the need for stakeholder involvement in an iterative process of design and evaluation.

KEYWORDS

large language models, 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, especially in high-stakes domains. In response, there has been extensive development of technical frameworks to formalize the concept of fairness so that it can be detected and enforced. Popular frameworks include fairness through unawareness, group fairness, and individual fairness [18]. The frameworks we have today are largely 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 [37], predicting recidivism in criminal justice [3], and text-based tasks such as coreference resolution [69].

Recently, there has been a surge of interest in generative AI, particularly the relatively general-purpose large language models (LLMs) that are trained for a foundational task such as next-word prediction, tuned with criteria such as conversational output or satisfying human preferences, and applied to many different use cases. These use cases span both traditional areas of concern for bias and fairness, such as evaluating resumes in hiring [5], and novel applications, such as drafting and editing emails [39], answering general knowledge queries [59], and code completion in software development [7].

In this paper, we consider whether and how the extant fairness frameworks can be applied to modern LLMs. We approach this mindful of both the hotly contested issues that already persist in the conventional fairness literature [e.g., 15] as well as the opportunities and challenges that other paradigms have presented, such as the difficulty in translating fairness frameworks to information access systems [21]. Our arguments are grounded in features of this new paradigm that seem essential for conceptualizing fairness.

First, at the algorithmic level, LLMs have exceptional flexibility. While their input and output have been largely restricted to natural language, recent history has shown that a wide range of content can be expressed in LLM-suitable natural language, and LLMs, or more broadly the class of so-called "foundation models" [11], are increasingly multimodal, such as the ability of GPT-4V [50] to take as input both natural language and images. 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, foregrounded in our analysis is the multitude of diverse stakeholders in LLM systems and their evolving relationships. As discussed in Section 3.2, the LLM is typically created and managed by a developer. The developer may curate the data the model uses, such as for training and retrieval, and they may also develop downstream LLM-based applications, though these may be done by other entities. As with other information systems, there are always users—whether individuals or groups who may have widely varying competencies [23]—and usually there are subjects of the content produced by the system, such as the people or groups being described in an information request. Additionally, there are researchers from academia, governments, nonprofits, or elsewhere aiming to better understand these systems and their societal impacts.

One important trend we see in the field today is that the model itself tends to be designed and trained by one entity with user-facing application development conducted by other groups or individuals. While it is good that multiple stakeholders participate in the LLM pipeline, this structure has led to information asymmetries that make it difficult to identify and mitigate harm. The sharing of information, primarily on the side of the LLM developers who

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

have information such as how the model was trained and what data it was trained on, and so on will be crucial for meaningful progress. With this information, researchers can probe capabilities and conduct particular context-based evaluations. This limited transparency was foregrounded in the February 2024 controversy in which the multimodal LLM Gemini developed by Google was found to diversify race and gender in images generated with prompts that specified historical settings that would be of a particular race and gender, such as soldiers and political figures in American and European historical settings that were almost exclusively men of European descent [48]. While there is much to be debated in how race and gender should be portrayed in image generation, third parties bemoaned the lack of information on the mechanisms by which these images were generated.

In what follows, Section 2 discusses work to date on LLM fairness, which focuses on association-based fairness metrics and practical challenges. Section 3 argues that in some cases there is a fundamental incompatibility between extant frameworks and modern LLM systems. Section 4 argues that, in the other cases, the flexibility of LLMs across data, use cases, stakeholders, and populations renders a general guarantee, stamp, or certificate of a fair LLM intractable. To conclude in Section 5, we articulate three principles motivated by these inherent challenges to move forward in the important goals of fairness and harm reduction in LLMs: the criticality of context, the responsibility of LLM developers, and the need for iterative and participatory design.

2 RECENT WORK ON LLM FAIRNESS

Transformer-based LLMs have been of great interest since they were introduced in 2017 [63], and this has accelerated since 2020 with the popularity of OpenAI's GPT models [49]. Many research teams have considered the role of algorithmic fairness in model evaluation, including a number of recent papers that have evaluated bias, discrimination, or fairness in the text generated by LLMs.

2.1 Association-based fairness metrics

Two recent reviews of this nascent literature [28, 42] enumerate a variety of fairness metrics, each of which constitutes an association between a feature of the embedding space or model output—word probabilities or generated text—and a sensitive attribute. This includes text measures a disparity of sentiment and toxicity in Wikipedia sentence completion across the profession, gender, race, religion, and political ideology of the article subject [16], the occurrence of violent words after a phrase such as “Two muslims walked into a” [2], and the topics brought up when completing sentences from fiction novels [46]. Other approaches include creating datasets of LLM continuations of text that stereotypes, demeans, or otherwise harms in ways related to gender and sexuality [27]; evaluations of conventional fairness measures when an LLM is used for a conventional machine learning task, such as predicting outcomes based on a text-converted tabular dataset [43]; recommending music or movies to a user who specifies their sensitive attribute such as race or religion [68]; and testing whether the model provides the same “yes” or “no” answer when asked for advice by a user who specifies their gender [61]. 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, disparity does not necessarily correspond to fairness in the sense predominant in the machine learning fairness literature or in other fields such as philosophy—as documented in Binns [6]. For example, in the framework of group fairness, which uses conditional equivalencies across sensitive attributes, mere disparity is known as demographic parity (see Definition 3), which is only one of many group fairness metrics and, while it is an important metric for comprehensive fairness evaluation and enforcement, achieving demographic parity is generally not viewed as achieving algorithmic fairness. In general, while the popular benchmarks such as WinoBias [69] and BBQ [52] that have been applied to LLMs capture important information about model behavior in relation to sensitive attributes, there is little reason to think that strong performance would imply fairness is achieved.

When existing work on LLMs has touched on richer notions of fairness, it has been in a highly constrained manner. For example, while Li et al. [42] briefly discussed counterfactual fairness (see Definition 7), they only do so by summarizing two papers that address it merely by perturbing the LLM input (e.g., converting Standard American English to African American English [44]), which does not acknowledge or address the fundamental challenges we present in Section 4.1 of how metrics fail to generalize across populations in which the data-generating process could vary significantly.

2.2 Practical challenges

Existing work has motivated and articulated significant challenges in evaluating and enforcing fairness in LLMs. Both Gallegos et al. [28] and Li et al. [42] summarize these, including the need to center marginalized communities through participatory design [8, 9] and develop better proxy metrics, such as to bridge the divide between intrinsic and extrinsic bias metrics [30]. These are important challenges to be addressed and remain so in light of the present work, but even if each of them were addressed, the fundamental challenges that are the focus of the present work would remain.

The fundamental challenges of LLM fairness have yet to be foregrounded in part because of the focus of existing work on relatively narrow use cases, often analyzing the LLM as a classifier or recommender in conventional machine learning use cases through the use of in-context learning to steer the model towards the conventional output format (e.g., a binary data label or recommendation) [43, 61, 68]. Given the flexibility of LLMs as text-to-text models, they can be deployed—though not necessarily with strong performance—to any conventional task in which the input and output can be expressed as a series of tokens. However, LLMs are not primarily used purely as substitutes for conventional, narrow-purpose models. The wide applicability of generated text has facilitated a wide range of applications, which are evolving every month as users and developers explore possibilities. Examples 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, gaming). 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

It is clear from the extant literature that achieving multiple fairness metrics simultaneously is intractable in most cases. Well-known impossibility results show that multiple group fairness metrics, such as those defined by rates of false positives and false negatives [14, 36] or demographic parity (Definition 3) and calibration (Definition 5) [36], cannot be simultaneously achieved in real-world environments. We argue, however, that challenges of LLM fairness run deeper in that some frameworks cannot even be applied.

3.1 Unawareness is impossible by design

Though often used as a strawman in the algorithmic fairness literature, perhaps the most common approach to fairness has been the default of fairness through unawareness (FTU)—simply leaving sensitive attributes out of the model's input.

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 largely emerged by considering models built on structured data, where data is organized into different variables that are used for prediction or classification. For example, a financial lending model could use a person's age, sex, and credit score to make a prediction about loan repayment. In this case, FTU could simply mean not using the field "sex" as an input to the predictive model. Although it has been established that simply omitting a sensitive field from a model at training time is insufficient to guarantee that no unwanted correlations with that attribute will exist in the model, there are still some cases where legal, policy, or feasibility constraints lead to this approach being used even though it is "widely considered naive" [66]. In one of the most widely known allegations of algorithmic discrimination, heterosexual spouses who used the Apple Card to make purchases noticed that the woman in the marriage was extended a much lower credit limit than the man. 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" [62].

The main critique of this approach is that the sensitive information is, often strongly, related to other features included in the model, so flexible models effectively recover the excluded sensitive attributes, which offsets potential fairness benefits.

By design, LLMs are trained on unstructured natural language, a context in which FTU is impossible because of the pervasiveness of sensitive attributes in natural language. Efforts to remove sensitive attributes may result in incoherence or distortion. 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, and other choices for how to remove national origin could result in particular distortions. Substituting a neutral phrase, "a country," could remove important narrative information, such as if the author intended to convey that Alice visited Brazil, the only South American country in which Portuguese is an official language.

Also, 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 the first-person being more common in groups of lower status [33]. In languages with gendered nouns (e.g., Spanish, German), enforcing gender fairness may require introducing entirely new vocabulary, and if nationality, native language, or other attributes of cultural background are considered sensitive, then languages, dialects, and subdialects seem extremely difficult if not impossible to extirpate from text—not to mention the entanglement of religion and other belief systems. It seems infeasible to enforce fairness with respect to all relevant sensitive attributes across large bodies of text while retaining sufficient information for model performance. There may also be direct ethical issues with the modification of text. Individual authors may not be okay with their text being modified, and even if the changes each seem to not change the author's intended meaning, it would be difficult to ever guarantee that no intended meaning was lost or even, at scale, whether factual information was edited.

As with many of the obstacles to fairness in LLMs—even if these conceptual challenges were addressed, the lack of transparency into modern LLMs would still make evaluating FTU impossible. 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 today—at least to third-party researchers and auditors. This is especially concerning from the FTU perspective. While conventional FTU explicitly leaves out the sensitive attribute, some approaches utilize 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 [45, 53]. The lack of LLM documentation prevents researchers and socially aware application developers from enforcing FTU, whether explicit or implicit, and 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 or multi-stakeholder fairness.

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 output (C-fairness), the providers of its output (P-fairness), the consumers and providers of its output considered together (CP-fairness), and the subjects of the items that are being provided (S-fairness).

Stakeholders are typically divided into the consumers, providers, and subjects of content in the system [1, 12, 21, 58]. For consumers—the people or organizations who receive the recommendations—there are many possible fairness targets [20, 22]; a common one is equity of utility, in which different users or user groups should receive comparably high-quality recommendations [24, 47, 64], which can also be applied to many LLM use cases.

For subjects—the people or organizations who are portrayed in the provided content—similar metrics may obtain and transfer straightforwardly to the LLM content. For example, early in the days of fairness in information access systems, it was noted that when searching for images of "CEO," Google returned a set of images largely depicting men with the first woman displayed being "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 in the LLM context. In general, it is not clear what distribution of representation should obtain in the system, and it is not clear how to estimate utility in order to achieve some distribution of utility at scale. 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 [26, 35, 55].

For providers—the people or organizations whose content is recommended—the 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 any case, fairness to providers is a matter of how the exposure of those providing content to the system is allocated to consumers.

In the use case of LLMs for information retrieval and management, this framework can at times transfer directly. For example, if someone searches for "coffee shops in Chicago" in an LLM chat or search interface, fairness could be defined in terms of equitable exposure to the different brick-and-mortar coffee shops in Chicago. 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 providers.

If users are searching for information in the LLM system, such as asking, "How are coffee beans roasted?" then LLMs can entirely circumvent the providers and upend the conventional notion of provider-side fairness. If the LLM extracts information from websites without directing users to the original source content, then it may be that none of the providers 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, which is taking all of the utility away from providers.

4 LLMS ARE TOO FLEXIBLE TO BE GENERALLY FAIR

Much of the excitement about LLMs is based on their flexibility across wide ranges of user input, tasks, contexts, and types of output. This has led to a characterization as "foundation models" [11]. In this sense, LLMs have become less like conventional models that perform specific tasks and have instead moved in the direction of large, automated repositories of information or like a human agent or crowdworker that can be assigned downstream tasks. This flexibility reduces the applicability of extant fairness frameworks to the LLM itself.

4.1 Group fairness doesn't generalize across populations

Group fairness metrics require an independence between model classification and sensitive attributes, often conditional on some relevant context such as the ground-truth labels that the model aims to predict. Three common metrics are:

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

Definition 4. (Equalized odds). A model achieves equalized odds if its predictions are statistically independent from the 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 from the 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 equal false positive rates (equalized odds), and equal precision (calibration). Recent work has also extended these notions to minimizing the maximum group error rate to help the worst-off group [17]. Conventionally, group fairness requires knowing the sensitive attributes, though recent work has also considered approaches for when the sensitive attributes are unavailable [34, 40, 71]. There are a number of methods for enforcing group fairness metrics, such as the pre-processing of datasets proposed by Feldman et al. [25] to guarantee bounds on demographic parity and the more recent method proposed by Johndrow and Lum [32] that works with a wider variety of datasets.

LLMs present a challenge for group fairness metrics because LLMs tend to be deployed across a wide range of data distributions. Lechner et al. [41] rigorously showed that it is impossible to build a non-trivial model that will perform fairly across all different data distributions, such as regions or demographic groups, to which it might be applied. This is a problem for fair regression and classification in general. For example, in recidivism prediction, fairness is assessed at a local level (e.g., counties in the U.S.) to ensure that the model is performing fairly and appropriately for that location's particular demographic mix and characteristics. However, this impossibility is especially problematic for LLMs because of the wide range of applications.

In general, it is not clear what an appropriate base population would be on which to detect and achieve group fairness. 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 the aforementioned Li and Zhang [43] do in predicting the label of a text-converted tabular dataset such as COMPAS. However, the LLM training data does not provide a clear base population because it is not a structured database comprised of 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 on which the model is trained.

Generalization across populations is also a concern for frameworks other than group fairness because of the wide range of textual, application, and social contexts at play in LLMs [56]. Here, we consider two examples: individual fairness [18] and counterfactual fairness, which is the most common causal notion of fairness [38].

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 so many different contexts—or if multiple metrics were applied, how these could be judiciously selected and guaranteed in every possible context for a single LLM. In terms of causal fairness, including counterfactual fairness, 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 in different populations. Again, these issues are not specific to LLMs, but they present substantial issues for the idea of a "stamp" or "certificate" of fairness for any model that is used in different populations, especially one with as little sense of a base population or an agreed-upon standard as modern LLMs.

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. Each of the different distributions and different use cases 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 fundamental to the popular framework of fair representations within the model or a representation produced by one model and utilized by another [67].

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 this framework, a machine learning system (e.g., classifier) first maps the dataset of individuals to a probability distribution in a new representation space, such that the system preserves as much information as possible about the individual while removing all information about the individual's protected class. The most well-known example of this approach is Bolukbasi et al. [10], which rigorously documented gender bias in Google News word embeddings, namely an association between occupations and a gender vector (i.e.., the dimension spanned from words like "men" on one end to words like "women" on the other), such that computer programmer was coded as highly male while homemaker was coded as highly female [10]. Indeed, this is where much of the NLP fairness literature has focused, documenting many similar biases across different word embedding models Sesari et al. [see 57, for a review].

That literature has developed debiasing approaches focused on the sensitive attribute dimension in the semantic space (e.g., $\vec{he} - \vec{she}$), such as zeroing the projection of each word vector (e.g., each occupation) onto the dimension itself [10] or training the model to align the gender dimension with the last coordinate, so that it can be easily removed or ignored [70]. However, Gonen and Goldberg [31] argue that such approaches "are mostly hiding the bias rather than removing it" because, even with the removal of such a dimension, word pairs tend to maintain their similarity, which still reflects associations with sensitive attributes—what Bolukbasi et al. [10] call "indirect bias."

In general, this presents a fundamental challenge for fairness in LLMs or other general-purpose systems because achieving fairness in one 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 use cases, 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. Attempts at debiasing for one context may remove or distort important information in another context.

The naive approach of debiasing the model with respect to the union of all potential sensitive attributes—even if it is 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 methods of narrowing a model to focus on specific tasks, such as fine-tuning, and currently we do not even have robust techniques to debias a single problematic relationship without incidentally obfuscating it or problematizing other relationships. The game of fairness whack-a-mole seems 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 use case, that would still require fine-grained adjustment to avoid counterproductive spillover into other areas.

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

Whether a model's behavior on a task is fair or desirable largely depends on how that output will be used. The output of one model is sometimes used as the input to another model, and fairness does not compose: Even if we can guarantee one task is fair, if that output is then plugged into another "fair" model, there is no guarantee the ultimate outcome will be fair [19]. For example, we previously referred to a popular benchmark dataset for assessing model bias, WinoBias, which presents a coreference resolution task in which models select one of two people to whom a pronoun most likely refers. Because these sentences involve different occupations (e.g., doctor, nurse), most models are found to be biased in the sense that they more readily associate the gendered pronoun with the occupation for which that gender is more prevalent [69]. However, even if a model that is determined to be fair by this metric or the benchmark itself were used in training a model, such as a different model to summarize the sentences, there is no guarantee on the fairness of the subsequent model's output.

However, composition may be able to address some of the challenges of fairness in LLMs. The general-purpose capabilities of deep neural networks could allow them to enforce 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 and, second, the instruction-tuning that allows them to obey natural language input. Many advances in LLM capabilities can be conceptualized as encouraging the model to improve its own output. For example, chain-of-thought prompting [65] encourages the model to first produce text that takes an incremental reasoning step towards its target. This can bolster performance by allowing the later token generations to build on the reasoning that the model has already generated, which has now 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 LLM. The most popular approaches, currently Reinforcement Learning from Human Feedback [RLHF; 51] and Direct Preference Optimization [DPO; 54], compel the model to steer itself towards human-provided preference data, and some other approaches, such as Constitutional AI [4] and `SELF-ALIGN` [60].

Model-assisted fairness strategies could take advantage of the power and flexibility of LLMs to address their own shortcomings. If the LLM itself has a sufficiently good internal representation of the biases at play, it could be prompted to impose the chosen fairness desiderata on its own output or that of another LLM. Of course, this creates a substantial risk of overreliance in doubling down on the blindspots of models, particularly those blindspots that are not yet sufficiently well-understood or appreciated, such that designers can guard against them. Recent approaches focus on model "self-correction." While there is skepticism that models can currently do this well, Ganguli et al. [29] 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, approaches like these that allow us to leverage those capabilities for fairness may allow us to make progress on the inevitably labyrinthine ethical challenges that future models will face.

5 PRINCIPLES

Over the past decade, the field of algorithmic fairness has developed multiple technical frameworks for assessing the impact of machine learning methods deployed in high-stakes domains. We argue that current frameworks are insufficient for the critical task of assessing the societal impact of deploying LLMs and do not believe it is feasible to certify that an LLM is "fair," in part because of inapplicability of some frameworks to LLM use cases, including unstructured natural language data, and in part because of the intractability of enforcing fairness on flexible, general-purpose foundational models, such as LLMs. Rather, mitigating harmful societal impacts from LLMs will require deeper engagement with the real-world tasks and contexts in which LLMs are ultimately used, as well as specific LLM practices such as prompt engineering and applying interpretability tools to the model.

For researchers evaluating societal impacts of LLMs, context is critical. A key strength of LLMs is that the same foundation model can be fine-tuned for an endless number of applications and 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 largely decontextualized, hypothetical, and ungrounded tests. 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. Proper contextualization has been lacking for years in the fairness literature [9], and the rise of LLMs increases its severity.

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 consequences. While users, regulators, researchers, and auditors have historically been well-positioned to collect and evaluate data in the later stages of this pipeline, there are substantial challenges in understanding and managing the earlier stages. 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 release far more information on actual usage and how the systems respond to real prompts from real users than is currently done [13]. LLM developers also have a responsibility to support these efforts through technical training, tools to facilitate evaluation of specific use cases, and other resources. Developers are still on the hook for mitigating harm—not downstream users and other stakeholders.

Managing the societal impact of LLMs will require iterative and participatory design and evaluation. Given the many different contexts in which these systems can be used, many with conflicting desiderata for defining fairness, we think it is impossible to make a generally "fair" LLM. Rather, LLM developers must work closely with third-party researchers, policymakers, end users, and other affected stakeholders in a participatory process that audits algorithms in the contexts in which they are used and mitigate any harmful effects, just as with any other widely deployed technology. Given the novel challenges of LLMs, the amplification of existing challenges, and the inevitable future developments, this process must also be iterative with frequent trials and assessments of new approaches to identifying and mitigating harm. While we are skeptical of many current approaches, there is still ample room and a strong imperative for responsible AI development and achieving fairness in particular use cases. In short, even though this problem is difficult, we believe addressing it is essential for responsible LLM development and deployment. With an iterative approach grounded in the nature of LLMs and real-world use, we believe that substantial progress could be made in a relatively short period of time.

REFERENCES

[1] 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* (March 2020), 127–158. https://doi.org/10.1007/s11257-019-09256-1

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

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

[4] 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]

[5] 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* (Sept. 2004), 991–1013. https://doi.org/10.1257/0002828042002561

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

[7] 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* (Dec. 2022), 35–57. https://doi.org/10.1145/3582083

[8] 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. https://doi.org/10.1145/3551624.3555290

[9] Su Lin Blodgett, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. 2020. Language (Technology) Is Power: A Critical Survey of "Bias" in NLP. https://doi.org/10.18653/v1/2020.acl-main.485

[10] 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]

[11] 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]

[12] Robin Burke. 2017. Multisided Fairness for Recommendation. arXiv:1707.00093 [cs]

[13] 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/.

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

[15] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. https://doi.org/10.1145/3097983.3098095

[16] 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. https://doi.org/10.1145/3442188.3445924 arXiv:2101.11718 [cs]

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

[18] 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]

[19] 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

[20] Michael D. Ekstrand, Lex Beattie, Maria Soledad Pera, and Henriette Cramer. 2024. Not Just Algorithms: Strategically Addressing Consumer Impacts in Information Retrieval.

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

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

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

[24] 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.

[25] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. https://doi.org/10.1145/2783258.2783311

[26] 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* (June 2022), 11882–11890. https://doi.org/10.1609/aaai.v36i11.21445

[27] 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. https://doi.org/10.18653/v1/2023.acl-long.343

[28] 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

[29] 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]

[30] 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

[31] 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. https://doi.org/10.18653/v1/N19-1061

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

[33] 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* (March 2014), 125–143. https://doi.org/10.1177/0261927X13502654

[34] 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

[35] Chen Karako and Putra Manggala. 2018. Using Image Fairness Representations in Diversity-Based Re-ranking for Recommendations. https://doi.org/10.1145/3213586.3226206

[36] 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

[37] 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. https://doi.org/10.1145/3514094.3534154

[38] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness.

[39] 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]

[40] 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]

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

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

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

[44] 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).

[45] Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. 2018. Does Mitigating MLs Impact Disparity Require Treatment Disparity?

[46] Li Lucy and David Bamman. 2021. Gender and Representation Bias in GPT-3 Generated Stories. https://doi.org/10.18653/v1/2021.nuse-1.5

[47] Rishabh Mehrotra, Ashton Anderson, Fernando Diaz, Amit Sharma, Hanna Wallach, and Emine Yilmaz. 2017. Auditing Search Engines for Differential Satisfaction Across Demographics. https://doi.org/10.1145/3041021.3054197

[48] Dan Milmo and Alex Hern. 2024. Google Chief Admits 'Biased' AI Tool's Photo Diversity Offended Users. *The Guardian* (Feb. 2024).

[49] OpenAI. 2022. Introducing ChatGPT.

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

[51] 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.

[52] 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]

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

[54] 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]

[55] 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]

[56] 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.

[57] Emeralda Sesari, Max Hort, and Federica Sarro. 2022. An Empirical Study on the Fairness of Pre-trained Word Embeddings. https://doi.org/10.18653/v1/2022.gebnlp-1.15

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

[59] 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]

[60] 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]

[61] 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]

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

[63] 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]

[64] Lequn Wang and Thorsten Joachims. 2021. User Fairness, Item Fairness, and Diversity for Rankings in Two-Sided Markets. https://doi.org/10.1145/3471158.3472260

[65] 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.

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

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

[68] 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. https://doi.org/10.1145/3604915.3608860 arXiv:2305.07609 [cs]

[69] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. https://doi.org/10.18653/v1/N18-2003

[70] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning Gender-Neutral Word Embeddings. https://doi.org/10.18653/v1/D18-1521

[71] Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. 2022. Towards Fair Classifiers Without Sensitive Attributes: Exploring Biases in Related Features. https://doi.org/10.1145/3488560.3498493 arXiv:2104.14537 [cs]