The Impossibility of Fair LLMs

Jacy Reese Anthis[1,2], Kristian Lum[1], Michael Ekstrand[3],
Avi Feller[4], Chenhao Tan[1]

[1]University of Chicago   [2]Stanford University
[3]Drexel University   [4]University of California, Berkeley

Abstract

The rise of general-purpose artificial intelligence (AI) systems, particularly large language models (LLMs), has raised pressing moral questions about how to reduce bias and ensure fairness at scale. Researchers have documented a sort of "bias" in the significant correlations between demographics (e.g., race, gender) in LLM prompts and responses, but it remains unclear how LLM fairness could be evaluated with more rigorous definitions, such as group fairness or fair representations. We analyze a variety of technical fairness frameworks and find inherent challenges in each that make the development of a fair LLM intractable. We show that each framework either does not logically extend to the general-purpose AI context or is infeasible in practice, primarily due to the large amounts of unstructured training data and the many potential combinations of human populations, use cases, and sensitive attributes. These inherent challenges would persist for general-purpose AI, including LLMs, even if empirical challenges, such as limited participatory input and limited measurement methods, were overcome. Nonetheless, fairness will remain an important type of model evaluation, and there are still promising research directions, particularly the development of standards for the responsibility of LLM developers, context-specific evaluations, and methods of iterative, participatory, and AI-assisted evaluation that could scale fairness across the diverse contexts of modern human-AI interaction.

1   Introduction

In response to the rapid adoption of machine learning systems and concerns about their negative societal impacts, researchers have developed compelling, nuanced technical frameworks to formalize ethical and social ideals—particularly the foundational notion of "fairness"—in order to systematically evaluate and apply them. Popular fairness frameworks include group fairness (Dwork et al.,

2011) and fair representations (Zemel et al., 2013). These frameworks have been extensively studied and applied to systems with structured data and specific use cases, such as the canonical examples of predicting default in financial lending (Kumar et al., 2022), predicting recidivism in criminal justice (Angwin et al., 2016), and coreference resolution in natural language (Zhao et al., 2018b).

There is an open question of how to think about bias fairness with the advent of generative AI and general-purpose large language models (LLMs). LLMs are increasingly used for a multitude of tasks that span both established areas of concern for bias and fairness—such as evaluating resumes in hiring, where the bias literature goes back decades (Bertrand and Mullainathan, 2004)—and areas less frequently discussed in the extant fairness literature—such as drafting and editing emails (Laban et al., 2023), answering general knowledge queries (Spatharioti et al., 2023), and software development (Bird et al., 2022).

We approach this topic mindful of both the hotly contested issues already present in the fairness literature (e.g., Corbett-Davies et al., 2017) and the challenges that other ascendant paradigms, such as information access systems (Ekstrand et al., 2022), have already presented for the ideal of fairness. For example, it is clear from the extant literature that multiple group fairness metrics, such as those defined by rates of false positives and false negatives (Chouldechova, 2017; Kleinberg et al., 2016) or demographic parity and calibration (Kleinberg et al., 2016), cannot be simultaneously achieved in real-world environments, even to an approximation.

We develop the stronger claim: fairness in the rigorous sense defined by these frameworks for narrow use, even on a single nontrivial metric, is intractable with general-purpose LLMs. The inherent challenges would persist regardless of advances in empirical methods, but we present future directions in light of them. Specifically, we make the

following arguments:

- Fairness through unawareness of sensitive attributes is made impossible by the unstructured training data and limited transparency of LLMs (Section 4.1).

- Standards for the fair treatment of content producers can be rendered obsolete by the LLM capacity for large-scale consumption and redistribution of content (Section 4.2).

- General-purpose LLMs cannot be made fair across many contexts because of the combinations of populations, use cases, and other factors that impose different fairness requirements (Section 4.3).

- Fairness does not compose, and LLM development and deployment involve the composition of different models, each with their own fairness challenges (Section 4.4).

- There is much important and tractable work to be done on LLM fairness, particularly in crafting standards of developer responsibility, refining in-depth methods for context-specific evaluation, and building scalable evaluations that iterate through participatory design and using AI capabilities to scale up to the multitude of real-world contexts (Section 5).

2 Approach

In order to assess the compatibility of LLMs with fairness frameworks, we considered each of the fundamental affordances of the LLM paradigm alongside each of the fairness frameworks. We see this as a broadly promising approach to examine new AI affordances across existing sociotechnical frameworks (e.g., agency (Sturgeon et al., 2025), deterrence theory (Hendrycks et al., 2025)).

First, at the technical level, we observe that LLMs have exceptional flexibility. It is increasingly clear that a wide range of content can be represented in LLM-suitable natural language. LLMs are increasingly multimodal, such as the capability of GPT-4 (OpenAI, 2023) to receive text, visual, audio, or mixed-modality input. LLMs lack the self-evident use case or even a relatively narrow set of use cases that have grounded prior work within these fairness frameworks. Recent work has demonstrated the need for metrics applicable to real-world deployment contexts and capable of

iterative refinement as systems evolve (Lum et al., 2024; Wallach et al., 2025; Weidinger et al., 2025).

Second, at the social level, our analysis foregrounds the multitude of diverse stakeholders in LLM systems and their continuously evolving relationships. As discussed in Section 4.2, there are developers: people and organizations who create datasets, curate datasets, develop models, deploy and manage models, and build downstream user-facing applications; there are users: subjects on which content produced by the system is based; and there are producers of content, such as owners of websites in the context of a search engine. In general, while our critiques are leveled at the applicability of technical frameworks, they echo the many challenges reported by practitioners from real-world deployment (Madaio et al., 2022a).

An illustrative problem that arises with many stakeholders is information asymmetry. Without information from developers (e.g., architecture details, training data), users and third parties have limited ability to conduct thorough evaluations. For a concrete example, consider the February 2024 public controversy in which Google's frontier LLM, Gemini, was found to diversify race and gender appearances in images even when prompts 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 (Milmo and Hern, 2024). 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, and the current LLM fairness literature does not fully address such complex cases with diverse stakeholders.

3 Recent work on LLM fairness

Interest in LLMs has accelerated in recent years as models such as ChatGPT, Claude, and Gemini have become more pervasive in everyday life, including sensitive contexts such as health and hiring. This has motivated research into many safety and ethical issues. While this paper is not intended as a comprehensive literature review, we first briefly review the recent work in machine learning and NLP research on bias and fairness in LLMs.

3.1 Association-based fairness metrics

Two recent reviews of this nascent literature (Gallegos et al., 2023; Li et al., 2024) enumerate a variety of fairness metrics that each constitute an association between a feature of the embedding space or model output and a sensitive attribute. NLP research in this area includes disparities of sentiment and toxicity in Wikipedia sentence completion across the profession, gender, race, religion, or political ideology of the article subject (Dhamala et al., 2021), the tendency to generate violent words after a phrase such as "Two muslims walked into a" [sic] (Abid et al., 2021), and variation in the topics introduced when completing sentences from fiction novels (Lucy and Bamman, 2021). Other approaches include creating datasets of LLM text continuations that include stereotypes, demeans, or otherwise harms in ways related to gender and sexuality (Fleisig et al., 2023); evaluating an LLM used for the conventional machine learning task of predicting outcomes based on a text-converted tabular dataset (Li et al., 2023); recommending music or movies to a user who specifies their sensitive attribute, such as race or religion (Zhang et al., 2023); and testing whether the model gives the same "yes" or "no" answer when asked for advice by users who specify their gender (Tamkin et al., 2023).

However, a lack of disparities in these test cases would not constitute fairness as conceptualized in technical frameworks or in other fields such as philosophy (e.g., Binns, 2021). For example, within the scope of group fairness, which uses conditional equivalencies of model output across sensitive attributes, the simplest notion—unconditional equivalence—is known as demographic parity. Demographic parity is an important metric to study, but achieving it (i.e., zero disparity) is rarely, if ever, viewed as achieving fairness. While the popular benchmarks that have been applied to LLM-generated text to date, such as WinoBias (Zhao et al., 2018b) and BBQ (Parrish et al., 2022), capture important information about the associations between generated text and sensitive attributes, strong model performance does not constitute fairness per se. Indeed, even without considering the technical fairness frameworks, the limitations of these benchmarks as proxies for issues such as stereotyping is well-established (Blodgett et al., 2021; Lum et al., 2024). There is little reason to think that the disparity measures, which are the most common fairness metrics in NLP, serve as

sufficient proxies for the fairness frameworks, even with narrow-purpose AI.

Extant work on LLMs has touched on the technical fairness frameworks, but that has typically been in a highly constrained manner. For example, while Li et al. (2024) briefly discussed counterfactual fairness, they only did so by summarizing two papers that merely perturb the LLM input, such as by converting Standard American English to African American English (Liang et al., 2023), which does not acknowledge or address the inherent challenges we present in Section 4.3 of how counterfactual fairness and other metrics fail to generalize across populations and how realistic counterfactuals would not merely vary in writing style or any other features directly observable in the text. Our work, in contrast, critiques the assumption that bias and fairness can be so easily measured.

3.2 Empirical challenges

Extant work has articulated significant challenges in achieving LLM fairness, but it has said little about the fairness frameworks that are used to measure and guarantee fairness in conventional machine learning and NLP applications. Gallegos et al. (2023) and Li et al. (2024) overview several issues, such as the need to center marginalized communities (Birhane et al., 2022; Blodgett et al., 2020) and to develop better proxies by bridging the divide between intrinsic and extrinsic bias metrics (Goldfarb-Tarrant et al., 2020). While we cannot presently cover all of the recent work on LLM fairness, including more recent reviews such as Chu et al. (2024b), we generally note that, even if every empirical challenge were addressed, the inherent challenges that are the focus of the present work would remain. We return to empirical challenges, and means to address them, in Section 5.

The inherent challenges of LLM fairness have yet to be foregrounded in part because work to date has largely focused on relatively narrow use cases. Often the LLM is applied as a classifier or recommender system in conventional machine learning tasks through the use of in-context learning to produce the conventional output format (e.g., a binary data label) (Li et al., 2023; Tamkin et al., 2023; Zhang et al., 2023). It is true that, given the flexibility of LLMs, they could be deployed to any conventional task, but LLMs are not primarily used or advertised as substitutes for conventional, narrow-purpose models. In the following enumeration of inherent challenges, we refer to various

studies that provide important conceptual foundations, but our claims are our own synthesis and not extracted directly from prior work.

4 Inherent challenges of fair LLMs

4.1 Unawareness is impossible by design

The framework of fairness through unawareness (FTU), which measures fairness based on whether the model input explicitly contains sensitive attributes, emerged for models built on structured data, typically in which data is organized into variables 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 in which FTU means that "gender" is excised from the training data. Legal, policy, and feasibility constraints often lead to the FTU approach in practice. In one of the most widely known allegations of algorithmic discrimination, a group of heterosexual married couples who used the Apple Card 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" (Telford, 2019).

By design, LLMs are trained on massive amounts of unstructured data, primarily natural language but also visual and audio modalities. FTU is impossible in these contexts because of the pervasiveness of sensitive attributes. Indeed, LLMs are readily able to infer personal characteristics such as the age, location, and gender of an author. For example, Staab et al. (2024) show that ChatGPT, Claude, and other LLMs can easily guess personal characteristics based on Reddit profiles.

Efforts to remove sensitive attributes can produce incoherence or distortion. For simplicity, we provide an example in which national origin (the sensitive attribute under consideration) is explicitly specified: Consider the sentence, "Alice grew up in Portugal, so Alice had an easy time on the trip to South America." Simply removing Alice's origin, "Portugal" or "in Portugal," would result in an ungrammatical sentence. Other approaches for removing national origin would still result in distortion. Substituting the neutral phrase "a country" or "in a country" would remove important narrative information, such as the author conveying that Alice visited Brazil, the only South American country

in which Portuguese is an official language. The story may go on to describe Alice's global travel, in which her national origin plays an important role in how she reacts to new experiences.

Efforts to remove more implicit sensitive attributes (e.g., of the text author) may result in even more distortion of content, and identifying them may be very challenging and has not been addressed in prior fairness studies (e.g., the aforementioned Liang et al. (2023)). Consider how relative status can be conveyed through pronoun usage, such as the use of first-person pronouns being more common in groups of lower social status (Kacewicz et al., 2014). Moreover, in languages with gendered nouns (e.g., Spanish, German), enforcing a notion of 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 authors not consenting to the modifications.

As with the other frameworks, FTU is additionally hindered by the current lack of model transparency. FTU would require that an LLM be documentably unaware of the sensitive information, which would require a level of documentation of training data that is unavailable for any state-of-the-art LLM today—at least to third-party researchers, auditors, and developers. Even with a model such as Llama, for which the weights are shared freely online, there is little public information about training data (Dubey et al., 2024). 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 given the strong predictive relationship (Lipton et al., 2018; Pope and Sydnor, 2011), which would be even more challenging.

4.2 Producer-side fairness criteria can be rendered obsolete

In the literature on fairness in recommender and information retrieval systems, the presence of multi-

ple stakeholders has motivated the multi-sided fairness framework. This framework requires that the system is fair with respect to each group of stakeholders, typically divided into consumers, subjects, and producers of content (Abdollahpouri et al., 2020; Burke, 2017; Ekstrand et al., 2022; Sonboli et al., 2022). For consumers and subjects (i.e., 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 (Ekstrand and Pera, 2022; Ekstrand et al., 2024; Mehrotra et al., 2017; Wang et al., 2021). While there are challenges in measuring quality or utility and what distribution of quality or utility is fair, these are more or less straightforwardly intensified from conventional NLP to LLM use cases.

For subjects, it may be difficult to define, detect, or enforce appropriate fairness metrics, particularly across modalities. For example, there is an open question of whether the target distribution of gender across search engine results for "CEO" 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 (Feng and Shah, 2022; Karako and Manggala, 2018; Raj and Ekstrand, 2022). These issues are compounded by the lack of clear correspondence between LLM outputs and real-world subjects: Images or texts produced by an LLM-based system often do not correspond directly to particular individuals or even particular levels of sensitive attributes, such as generating images that do not clearly represent a particular race or ethnicity. Note that we consider an image-producing system to still be an LLM given that natural language (e.g., English) is still the primary modality and "language" itself can be perceived more broadly to include the encoding and communication of ideas through imagery and other modalities.

There are more complex challenges in multi-sided fairness for producers, also known as providers. The conventional fairness target is typically an 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. This framework can at times transfer directly to LLMs in the context of information retrieval and management tasks. For example, if someone searches for "coffee shops in San Francisco" in an LLM chat—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 visit the cafes, which provides utility—fairly or unfairly—to the producers.

However, if users search for information via 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 utility from the producers and renders the conventional producer-side fairness criteria obsolete. This is a particularly important consideration given the ongoing integration of LLMs into search engines, such as OpenAI's SearchGPT (OpenAI, 2024) and Google. While these developers have committed to responsible practices, such as supporting content producers, third-party evaluation can help ensure accountability.

4.3 General-purpose LLMs cannot be made fair across many contexts

Much of the excitement surrounding LLMs is based on their general-purpose flexibility across wide ranges of populations, use cases, and sensitive attributes. This flexibility makes many conventional fairness metrics intractable, which we illustrate with the group fairness framework.

Group fairness metrics, such as demographic parity, equalized odds, and calibration (Verma and Rubin, 2018), 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). 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), or equal precision (calibration). Recent work includes

extensions of these notions, such as prioritizing the worst-off group (Diana et al., 2021); methods to estimate the sensitive attribute when it is unavailable (Kallus et al., 2021; Zhao et al., 2022; Lahoti et al., 2020); and methods of enforcement, such as preprocessing (Feldman et al., 2015).

4.3.1 Populations and use cases

LLMs, particularly general-purpose LLMs, present a challenge for group fairness metrics in part because LLMs tend to be deployed across a wide range of input and output distributions. Lechner et al. (2021) 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 prediction in criminal justice), fairness is typically targeted at a local jurisdiction, which ensures that the model is performing fairly for that location's particular demographic mix (e.g., age, race) but typically cannot also ensure fairness in different locations. The purpose and use of LLMs makes it infeasible to restrict them to this sort of target population. Interviews with AI practitioners have shown that this complexity is already a substantial challenge in the deployment of conventional AI systems (Madaio et al., 2022a).

In general, it is not clear what an appropriate target 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 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 et al. (2023) 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 target population because a corpus of text is not a structured database comprising people and their individual characteristics. An LLM may be trained in part on structured databases, but the output of the model for any such predictions is also based on the wide scope of unstructured training data. This is compounded when the LLM is deployed across many use cases within each population.

Generalization across populations and use cases 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 LLM use (Rauh et al., 2022). For two examples: First, individual fairness requires that the model output is Lipschitz continuous with respect to the model input (Dwork et al., 2011). In this case, it is not clear what similarity metrics could be reasonably applied across the multitude of populations or use cases. If context-specific metrics were applied, it is still left undetermined how these could be judiciously selected and guaranteed.

Second, counterfactual fairness requires that the model would have produced the same output for an individual if they had a different level of the sensitive attribute (Kusner et al., 2017). However, it is often difficult to identify the causal structure of the data-generating process in even a single 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 and use cases.

4.3.2 Sensitive attributes

Given the issues discussed in Section 4.1, it may not be tractable to exclude sensitive attributes from training data, and each of the populations and use cases can require fairness metrics to be enforced for a different set of sensitive attributes. The effort required increases combinatorially with the importance of intersections of sensitive attributes (Himmelreich et al., 2024).

This is a challenge for the group fairness metrics already defined, but the issue is particularly salient for the popular ideal of fair representations, which requires that data representations do not contain information that can identify the sensitive attributes of individuals (Zemel et al., 2013).

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. (2016), 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 (see Sesari et al., 2022, for a review of more recent work).

Researchers have developed a number of debiasing approaches focused on a particular sensitive

attribute dimension, such as zeroing the projection of each word vector onto the dimension (Bolukbasi et al., 2016) or training the model to align the sensitive attribute dimension with a coordinate of the embedding space so that it can be easily removed or ignored (Zhao et al., 2018a). However, Gonen and Goldberg (2019) showed that such approaches "are mostly hiding the bias rather than removing it" because word pairs tend to maintain similarity, reflecting associations with sensitive attributes in what Bolukbasi et al. (2016) call "indirect bias."

Achieving fairness in one LLM context may be contingent on alteration of the statistical relationships between the context-specific sensitive attribute and other features of the data, particularly the removal of information. 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 closely associated with gender, such as biological sex and pregnancy status, may be essential factors in medical decision-making. In general, attempts to debias 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, even for only its most important sensitive attributes, 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. This is infeasible with current methods, such as supervised 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. The game of fairness whack-a-mole appears intractable, dashing hopes of cross-context debiasing.

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

Whether a model's behavior is fair or desirable largely depends on how the model's output will be used. In many modern AI systems, the output of one model is often used as the input to another model, but this process—known as "composition"—is difficult 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 (2019). Ensuring fairness is particularly challenging when the different systems—such as OpenAI's ChatGPT (OpenAI, 2022) and DALL-E, OpenAI's primary text-to-image model (Ramesh et al., 2021)—operate with different modalities or training data. In the case of Google's Gemini model, the aforementioned February 2024 controversy was compounded by the difficulty of identifying how the text input was related to the image output (Milmo and Hern, 2024).

However, it may be possible to use the aforementioned flexibility of general-purpose LLMs to 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 (Wei et al., 2022) 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.

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; Ouyang et al., 2022) and Direct Preference Optimization (DPO; Rafailov et al., 2023), compel the model towards human-provided preference data, and some other approaches, such as constitutional AI (Bai et al., 2022) and SELF-ALIGN (Sun et al., 2023), 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." There is skepticism that models can currently do this well, but Ganguli et al. (2023) show impressive results on bias and discrimination benchmarks "simply by instructing models to avoid harmful outputs."

5 Implications and future research

We conclude with a brief discussion of how to move forward with building fair AI systems and researching LLM fairness in light of these challenges.

5.1 Developer responsibility

Fairness issues manifest throughout the end-to-end pipelines from AI model design and training to model deployment and long-term effects. Users, regulators, researchers, and auditors have historically been well-positioned to evaluate the later stages of this pipeline, but there are substantial challenges for their efforts to understand the earlier stages, hindering efforts towards goals such as fairness through unawareness is infeasible (Section 4.1). LLM developers have a responsibility to support users and third parties. For researchers and other third parties to conduct grounded evaluations, companies that deploy LLMs should share information on actual usage and how the systems respond to real prompts from real users (Caliskan and Lum, 2024; Lum et al., 2024). The challenges of unstructured data, producer equity, and diverse contexts suggest a need for LLM developers to work closely with third-party researchers, policymakers, end users, and other affected stakeholders in a participatory and context-informed design process (Muller and Kuhn, 1993).

Modern generative AI systems are trained with unprecedented amounts of natural language and multimodal data. In addition to lacking transparency of training data, the extensive data scraping efforts raise concerns about copyright and intellectual property law (Abbott and Rothman, 2023; Chu et al., 2024a). If a user or advertiser pays a search engine, that could be unfairly extracting value from both producers (as discussed in Section 4.2) of the search result content as well as producers of training data for the underlying LLM. The tendency of general-purpose LLMs to intake extremely large

training datasets also raises concerns about the filters used for selection, such as "quality" text filters that may disproportionately exclude certain voices (Lucy et al., 2024). Transparency challenges are compounded by the lack of evaluation infrastructure for LLMs, unlike transportation, aerospace, pharmaceuticals and other fields with mature evaluation regimes established through decades of institutional investment (Weidinger et al., 2025).

5.2 Context-specific evaluations

Building better general-purpose AI systems and measuring their fairness—even if we cannot say that the system is generally fair—will require articulating specific connections to real use cases and corresponding harms, adapting technical frameworks to the specificity of a particular context (e.g., Anthis and Veitch, 2023; Blandin and Kash, 2024). With the challenges of translating and composing fairness across models and contexts (Section 4.3), it is unlikely that any "trick tests," such as coreference resolution of gendered pronouns, will provide satisfactory evidence for or against LLM fairness (Lum et al., 2024). There has been a dearth of proper contextualization in the fairness literature (e.g., Aler Tubella et al., 2023; Blodgett et al., 2020), and the intractability of generalized fairness adds weight to this critique.

Bias is often present from pretraining data, such as large-scale internet corpora, but the interactive feedback loops of LLM prompt engineering, custom instructions, and supervised fine-tuning risk amplifying biases by further shifting the context in which the LLM operates. Users who speak low-resource languages already face lower model performance (e.g., OpenAI et al., 2024), a challenge that can be compounded by a limited ability to iterate on prompting strategies. As a user continues to interact with a system, even small biases can be amplified, and guardrails can erode—though empirical research is needed on how this manifests with modern LLM use. As LLMs create more content and synthetic data that is used in training new systems, they can exacerbate the deterioration of public goods, including "enshittification" as people become locked into online platforms and content quality deteriorates (Doctorow, 2025).

In 2024, the risks of LLM extensibility and customization became salient in public policy debates, such as the European Union AI Act and California's SB 1047, the *Safe and Secure Innovation for Frontier Artificial Intelligence Models Act*, which

would ostensibly require that LLM developers implement safety guarantees that include the prevention of misuse and the capability to promptly shut down the system if necessary. Critics have argued that LLM developers cannot make such guarantees because LLMs are inevitably deployed in new and unexpected contexts and are able to be substantially modified by users and third parties, particularly with the development of open source models that allow academics and independent developers to research and innovate.

5.3 Scalable evaluation

Today, developing fairness metrics for a single context requires substantial effort to iteratively study harms and develop mitigations. The difficulty increases combinatorially with the variety of populations, use cases, and sensitive attributes and their intersections (Himmelreich et al., 2024), across which any realistic amount of effort is insufficient. Intensive strategies that interview and co-design with stakeholders can supply clarity (Madaio et al., 2022b), but ideally there would be more scalable evaluations that—while LLM fairness guarantees are intractable—can meaningfully support fairness across the many different LLM use cases.

We believe there is an exciting and largely untapped opportunity at the intersection of technical fairness frameworks (and related technical frameworks, such as privacy) and scalable human-AI alignment. For example, expanding on our argument in Section 4.4, the ethics-driven methodologies of RLHF (Ouyang et al., 2022), DPO (Rafailov et al., 2023), constitutional AI (Bai et al., 2022) and SELF-ALIGN (Sun et al., 2023) can each incorporate technical frameworks and context-adaptive methods. This can include feeding contextual information to a model that meaningfully synthesizes it in a human-like way and adjusts accordingly. There can also be AI-automated evaluation pipelines, including "LLM-as-a-judge" (Kanepajs et al., 2025; Zheng et al., 2023) fairness rubrics, high-quality simulations of human data (Anthis et al., 2025), and the generation and validation of tests or simulated user queries (Sturgeon et al., 2025) that probe fairness at scale.

One of the primary challenges of this research direction will be accounting for bias from each of the inputs, such as the reward model in RLHF that inevitably comes from particular people, raising questions about to whose values LLMs are being aligned (Gabriel, 2020); which of their values

LLMs are being aligned, such as whether ratings are based on "helpfulness," "honesty," or "harmlessness" (Askell et al., 2021; Liu et al., 2024); and at what times that input is elicited, given changes in values over time (Carroll et al., 2024). One could also utilize interpretability tools (e.g., Singh et al., 2023; Nanda and Bloom, 2022) that allow participatory and iterative exploration of how bias manifests. Again, bias can manifest through these tools, such as the quality of interpretation provided to different users in different contexts or variation in the tendency of LLMs to refuse user requests across groups (Wester et al., 2024). On the other hand, if interpretability tools succeed in providing relevant information, this can be used to make fairer models as they allow for more context-specific adjustments. The potential for harm and benefit will depend in part on the quality of future interpretability tools, especially because RLHF and related techniques tend to incorporate myopic and biased input (Casper et al., 2023), which could lead to overconfident or otherwise inaccurate explanations.

It remains true that humans cannot feasibly scale manual efforts to the scope of LLMs despite the risk of "bias all the way down" when using AI tools to address AI issues. Moreover, AI tools have the unique advantage that their capabilities will scale alongside the risks from powerful AI systems. AI tools should be considered across not just fairness but the host of AI issues (e.g., privacy, mental health) to make progress towards safe and beneficial general-purpose AI.

6 Conclusion

Work to date has measured associations and disparities as LLMs provide different output when different demographics are specified explicitly (e.g., "White" and "Black") or implicitly (e.g., dialects). However, these cannot substitute for more rigorously developed fairness frameworks, such as group fairness and causal fairness. When we consider these frameworks, the inherent challenges render general fairness impossible for an LLM. While these limits are underappreciated in the current literature, there are promising research and practical directions in standards of developer responsibility, context-specific evaluations, and scaling evaluation in ways that cautiously utilize general-purpose AI while mitigating the amplification of moral issues within those systems as well.

Limitations

Given the complexity and opacity of today's deep neural networks, it is difficult to formally analyze their capabilities and limitations. The preceding claims and analysis were not developed mathematically, and the outcome of such analysis would depend on particular assumptions and operationalizations. It is also possible that compelling new technical frameworks, perhaps developed specifically for general-purpose LLMs or other general-purpose systems, will circumvent the inherent challenges we described. Finally, while we believe it is important to lay out a conceptual foundation of what is and is not possible, there are many open empirical challenges that we have not addressed in this work, particularly the quantification of how much fairness metrics can be partially satisfied in real-world settings and the development of scalable methods for context-specific alignment with fairness and other social values.

Acknowledgments

References

Ryan Abbott and Elizabeth Rothman. 2023. Disrupting creativity: Copyright law in the age of generative artificial intelligence. *Florida Law Review*, 75(6):1141–1201.

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):127–158.

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Large language models associate Muslims with violence. *Nature Machine Intelligence*, 3(6):461–463.

Andrea Aler Tubella, Dimitri Coelho Mollo, Adam Dahlgren Lindström, Hannah Devinney, Virginia Dignum, Petter Ericson, Anna Jonsson, Timotheus Kampik, Tom Lenaerts, Julian Alfredo Mendez, and Juan Carlos Nieves. 2023. ACROCPoLis: A Descriptive Framework for Making Sense of Fairness. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1014–1025, Chicago IL USA. ACM.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *ProPublica*.

Jacy Reese Anthis, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. 2025. LLM Social Simulations Are a Promising Research Method. *arXiv preprint*. ArXiv:2504.02234 [cs].

Jacy Reese Anthis and Victor Veitch. 2023. Causal context connects counterfactual fairness to robust prediction and group fairness. In *Advances in Neural Information Processing Systems*, volume 36, pages 34122–34138. Curran Associates, Inc.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A General Language Assistant as a Laboratory for Alignment. *Preprint*, arXiv:2112.00861.

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. *Preprint*, arXiv:2212.08073.

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):991–1013.

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

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):35–57.

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*, pages 1–8, Arlington VA USA. ACM.

Jack Blandin and Ian A. Kash. 2024. Learning Fairness from Demonstrations via Inverse Reinforcement Learning. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 51–61, Rio de Janeiro Brazil. ACM.

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*, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

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]*.

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

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/.

Micah Carroll, Davis Foote, Anand Siththaranjan, Stuart Russell, and Anca Dragan. 2024. AI Alignment with Changing and Influenceable Reward Functions. *Preprint*, arXiv:2405.17713.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. 2023. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research*.

Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2):153–163.

Timothy Chu, Zhao Song, and Chiwun Yang. 2024a. How to Protect Copyright Data in Optimization of Large Language Models? *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17871–17879.

Zhibo Chu, Zichong Wang, and Wenbin Zhang. 2024b. Fairness in Large Language Models: A Taxonomic Survey. *Preprint*, arXiv:2404.01349.

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*, pages 797–806, Halifax NS Canada. ACM.

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

Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. 2021. Minimax Group Fairness: Algorithms and Experiments. *Preprint*, arXiv:2011.03108.

Cory Doctorow. 2025. *Enshittification: Why Everything Suddenly Got Worse and What to Do About It*. Farrar, Straus and Giroux.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu,

Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich,

Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal

Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

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

Cynthia Dwork and Christina Ilvento. 2019. Fairness Under Composition. *Leibniz International Proceedings in Informatics*, pages 20 pages, 627743 bytes.

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*.

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):1–177.

Michael D. Ekstrand and Maria Soledad Pera. 2022. Matching Consumer Fairness Objectives & Strategies for RecSys. *Preprint*, arXiv:2209.02662.

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*, pages 259–268, Sydney NSW Australia. ACM.

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):11882–11890.

Eve Fleisig, Aubrie Amstutz, Chad Atalla, Su Lin Blodgett, Hal Daumé Iii, Alexandra Olteanu, Emily Sheng, Dan Vann, and Hanna Wallach. 2023. Fair-Prism: Evaluating Fairness-Related Harms in Text Generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6231–6251, Toronto, Canada. Association for Computational Linguistics.

Iason Gabriel. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3):411–437.

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*.

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. *Preprint*, arXiv:2302.07459.

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*.

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

Dan Hendrycks, Eric Schmidt, and Alexandr Wang. 2025. Superintelligence Strategy: Expert Version. *arXiv preprint*. ArXiv:2503.05628 [cs].

Johannes Himmelreich, Arbie Hsu, Kristian Lum, and Ellen Veomett. 2024. The Intersectionality Problem for Algorithmic Fairness. *arXiv preprint*. ArXiv:2411.02569 [cs].

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):125–143.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. 2021. Assessing Algorithmic Fairness with Unobserved Protected Class Using Data Combination. *Management Science*.

Arturs Kanepajs, Aditi Basu, Sankalpa Ghose, Constance Li, Akshat Mehta, Ronak Mehta, Samuel David Tucker-Davis, Eric Zhou, and Bob Fischer. 2025. What do Large Language Models Say About Animals? Investigating Risks of Animal Harm in Generated Text. *arXiv preprint*. ArXiv:2503.04804 [cs].

Chen Karako and Putra Manggala. 2018. Using Image Fairness Representations in Diversity-Based Reranking for Recommendations. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*, pages 23–28, Singapore Singapore. ACM.

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.*

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*, pages 357–368. ACM.

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

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

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. *Preprint*, arXiv:2006.13114.

Tosca Lechner, Shai Ben-David, Sushant Agarwal, and Nivasini Ananthakrishnan. 2021. Impossibility results for fair representations. *Preprint*, arXiv:2107.03483.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2024. A survey on fairness in large language models. *Procedia Computer Science*, 00:1–28.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks Are All You Need II: Phi-1.5 technical report. *Preprint*, arXiv:2309.05463.

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.*

Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. 2018. Does mitigating MLs impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Ryan Liu, Theodore Sumers, Ishita Dasgupta, and Thomas L. Griffiths. 2024. How do Large Language Models Navigate Conflicts between Honesty and Helpfulness? In *Forty-First International Conference on Machine Learning.*

Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

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. *Preprint*, arXiv:2401.06408.

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

Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022a. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–26.

Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022b. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–26.

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*, pages 626–633, Perth, Australia. ACM Press.

Dan Milmo and Alex Hern. 2024. Google chief admits 'biased' AI tool's photo diversity offended users. *The Guardian.*

Michael J. Muller and Sarah Kuhn. 1993. Participatory design. *Communications of the ACM*, 36(6):24–28.

Neel Nanda and Joseph Bloom. 2022. TransformerLens.

OpenAI. 2022. Introducing ChatGPT.

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

OpenAI. 2024. SearchGPT Prototype. https://openai.com/index/searchgpt-prototype/.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin,

Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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. *Preprint*, arXiv:2110.08193.

Devin G Pope and Justin R Sydnor. 2011. Implementing Anti-Discrimination Policies in Statistical Profiling Models. *American Economic Journal: Economic Policy*, 3(3):206–231.

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. *Preprint*, arXiv:2305.18290.

Amifa Raj and Michael D. Ekstrand. 2022. Fire Dragon and Unicorn Princess; Gender Stereotypes and Children's Products in Search Engine Responses. *Preprint*, arXiv:2206.13747.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. *Preprint*, arXiv:2102.12092.

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*, volume 35, pages 24720–24739. Curran Associates, Inc.

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)*, pages 129–144, Seattle, Washington. Association for Computational Linguistics.

Chandan Singh, Aliyah R. Hsu, Richard Antonello, Shailee Jain, Alexander G. Huth, Bin Yu, and Jianfeng Gao. 2023. Explaining black box text modules in natural language with language models. *Preprint*, arXiv:2305.09863.

Nasim Sonboli, Robin Burke, Michael Ekstrand, and Rishabh Mehrotra. 2022. The multisided complexity of fairness in recommender systems. *AI Magazine*, 43(2):164–176.

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. *Preprint*, arXiv:2307.03744.

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.

Benjamin Sturgeon, Leo Hyams, Daniel Samuelson, Ethan Vorster, Jacob Haimes, and Jacy Reese Anthis. 2025. HumanAgencyBench: Do Language Models Support Human Agency?

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. *Preprint*, arXiv:2305.03047.

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. *Preprint*, arXiv:2312.03689.

Taylor Telford. 2019. Apple Card algorithm sparks gender bias allegations against Goldman Sachs. *Washington Post*.

Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. In *Proceedings of the International Workshop on Software Fairness*, pages 1–7, Gothenburg Sweden. ACM.

Hanna Wallach, Meera Desai, A. Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Nicholas Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. 2025. Position: Evaluating generative ai systems is a social science measurement challenge. *Preprint*, arXiv:2502.00561.

Jindong Wang, Wenjie Feng, Chang Liu, Chaohui Yu, Mingxuan Du, Renjun Xu, Tao Qin, and Tie-Yan Liu. 2021. Learning Invariant Representations across Domains and Tasks. *arXiv:2103.05114 [cs, eess]*.

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*, volume 35, pages 24824–24837. Curran Associates, Inc.

Laura Weidinger, Inioluwa Deborah Raji, Hanna Wallach, Margaret Mitchell, Angelina Wang, Olawale Salaudeen, Rishi Bommasani, Deep Ganguli, Sanmi Koyejo, and William Isaac. 2025. Toward an evaluation science for generative ai systems. *Preprint*, arXiv:2503.05336.

Joel Wester, Tim Schrills, Henning Pohl, and Niels Van Berkel. 2024. "As an AI language model, I cannot": Investigating LLM Denials of User Requests. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–14, Honolulu HI USA. ACM.

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

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*, pages 993–999.

Han Zhao, Chen Dan, Bryon Aragam, Tommi S. Jaakkola, Geoffrey J. Gordon, and Pradeep Ravikumar. 2022. Fundamental limits and tradeoffs in invariant representation learning. *Journal of Machine Learning Research*, 23(340):1–49.

Han Zhao, Shanghang Zhang, Guanhang Wu, José M. F. Moura, Joao P Costeira, and Geoffrey J Gordon. 2018a. Adversarial Multiple Source Domain Adaptation. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018b. Gender bias in coreference resolution: Evaluation and debiasing methods. *Preprint*, arXiv:1804.06876.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.