

Bias Is a Math Problem, AI Bias Is a Technical Problem: 10-Year Literature Review of AI/LLM Bias Research Reveals Narrow [Gender-Centric] Conceptions of ‘Bias’, and Academia-Industry Gap

Sourojit Ghosh, Kyra Wilson

University of Washington, Seattle
 {ghosh100, kywi}@uw.edu

Abstract

The rapid development of AI tools and implementation of LLMs within downstream tasks has been paralleled by a surge in research exploring how the outputs of such AI/LLM systems embed biases, a research topic which was already being extensively explored before the era of ChatGPT. Given the high volume of research around the biases within the outputs of AI systems and LLMs, it is imperative to conduct systematic literature reviews to document throughlines within such research. In this paper, we conduct such a review of research covering AI/LLM bias in four premier venues/organizations – *ACL, FAccT, NeurIPS, and AAI – published over the past 10 years. Through a coverage of 189 papers, we uncover patterns of bias research and along what axes of human identity they commonly focus. The first emergent pattern within the corpus was that 155/189 papers did not establish a working definition of ‘bias’ for their purposes, opting instead to simply state that biases and stereotypes exist that can have harmful downstream effects while establishing only mathematical and technical definition of bias. 94 of these 155 papers have been published in the past 5 years, after literature reviews were published with a similar finding about NLP research and recommendation to consider how such researchers should conceptualize bias, going beyond strictly technical definitions. Furthermore, we find that a large majority of papers – 151/189 papers – focus on gender bias (mostly, gender and occupation bias) within the outputs of AI systems and LLMs. By demonstrating a strong focus within the field on gender, race/ethnicity (57/189 papers), age (39/189 papers), religion (36/189 papers) and nationality (25/189 papers) bias, we document how researchers adopt a fairly narrow conception of AI bias by overlooking several non-Western communities in fairness research, as we advocate for a stronger coverage of such populations. Finally, we note that while our corpus contains several examples of innovative debiasing methods across the aforementioned aspects of human identity, only 20/189 papers include recommendations for how to implement their findings or contributions in real-world AI systems or design processes. This indicates a concerning academia-industry gap, especially since many of the biases that our corpus contains several successful mitigation methods that still persist within the outputs of AI systems and LLMs commonly used today. We conclude with recommendations towards future AI/LLM fairness research, with stronger focus on diverse marginalized populations.

1 Introduction

Though Artificial Intelligence (AI) is by no means a technology novel to the 21st century, the development of the AI assistant ChatGPT powered by the Large Language Model (LLM) GPT-3 and its subsequent release for public usage in November 2022 spurred an era of promise and heralded the still-ongoing AI summer. As billions of dollars are continually poured across the world into the design and adoption of novel AI systems and tools leveraging multimodal LLMs in downstream tasks alongside the promulgation of new policies and legislations such as Executive Order 14179: *Removing Barriers to American Leadership in Artificial Intelligence* (2025) in the US, there has also been a steep rise in LLM-related research throughout this period of time (Pang et al. 2025) towards technical innovation and expanding the scope of tasks which AI systems and LLMs can accomplish.

In addition, a significant proportion of this research has focused on the various ways in which the outputs of AI systems and LLMs exhibit bias towards traditionally marginalized populations (Gupta et al. 2023; Wan et al. 2024). This is a critical angle of research exploring the sociotechnicality of AI systems and LLM-powered tools, recognizing that the presence of negative stereotypes within their outputs against populations historically marginalized in society and oppressed through existing societal structures brings further harm upon them and exacerbates their marginalization. However, past research both in and adjacent to AI/LLMs, namely that of natural language processing (NLP), has demonstrated that researchers focusing on the nuances of biases within system outputs commonly display purely technical understandings of “bias” in their research (e.g, Blodgett et al. 2020). The combination of the rapid publication of AI/LLM bias research (Pang et al. 2025) and a previously-recognized narrow conception of “bias” by technical researchers in related fields necessitates an exploration of the body of work in this field.

In this paper, we perform a systematic literature review covering work published in four premier venues/organizations — *ACL, FAT*/FAccT, NIPS/NeurIPS, and AAI — exploring emergent trends within bias and fairness research in AI systems and LLMs published within the last 10 years. Our coverage includes 189 full papers, curated through computational and manual screening, as we make the following three contributions to the field:

1. Firstly, we observe that 82% (155/189) papers studying “bias” in AI systems or LLMs do not actually define “bias”, and 74 of these 136 have been published in the last 5 years during the massive surge of LLM/AI fairness research. This is particularly concerning in light of Blodgett et al. (2020)’s similar finding 5 years ago, as it is indicative that researchers in the field have not since taken their call to heart. The lack of a definition of “bias” in a paper studying “bias” poses real problems, and we provide recommendations for researchers to adopt more than mathematical conceptions of bias in their work.
2. We document a strong focus on ‘gender bias’ within the set of papers surveyed within our dataset, with 79.9% (151/189) papers demonstrating such a focus. Alongside a focus on gender (most prominently, the associations of gender with occupations, as in 19.6% or 37/189 papers), 30.2% (57/189) researchers also studied race/ethnicity bias (often equating the two), 13.2% (25/189) studied nationality bias, 20.6% (39/189) focused on age bias, and 19.1% (36/189) researchers worked on religion bias. Our findings highlight a concerning trend within the field that research around AI bias and fairness is often restricted to axes of identity perceived globally and interpreted through Western lenses (e.g., gender and occupation bias research often focuses based on US occupations data, examples of racial bias in word embeddings refers to African-Americans as criminals, etc.). This pattern has been alluded to in the past (e.g., Dev et al. 2023; Ghosh 2024; Qadri et al. 2023), but not explicitly demonstrated through such a coverage of the field as ours. We highlight potential avenues to pursue research around historically marginalized communities often overlooked in AI-fairness research, to achieve stronger coverage of users and communities producing or consuming AI outputs.
3. Finally, it is evident across the corpus of papers that though researchers excel at innovative and effective debiasing techniques across the aforementioned axes of identity and providing technical solutions towards fairer AI systems, only 10.6% (20/189) papers contain any actionable implementation techniques of their findings within real-world AI systems. This is highlighted by an absence of potential design recommendations or mention of implementation procedures in papers, beyond making codebases and datasets publicly available for academic transparency. This finding is particularly salient because it reflects a glaring academia-industry gap: at the same time as a large volume of research is being published that is producing several methods to mitigate different biases which are demonstrated to be highly successful, real-world AI systems and LLMs commonly used in downstream tasks contain and perpetuate such biases within their outputs. We call for stronger collaboration between academic researchers and industry developers of AI systems, and advocate for academics to include within their writing strategies for the applications of their work within commonly-used AI systems.

2 Background

2.1 Literature Reviews

A literature review is a common and effective way of studying the state of a field and the various patterns of research within it, allowing for a collective understanding of salient threads of research and the identification of gaps to be filled with further research (Knopf 2006). Literature reviews are commonly of two types – systematic literature reviews that meticulously follow replicable protocols for screening and evaluation (Kitchenham 2004), and semi-systematic literature reviews that afford more flexibility and subjectivity within curation and review processes (Snyder 2019). A few recent examples include Mack et al. (2021)’s work exploring the field of Accessibility research provided a discursive overview of the field in terms of preferred communities of focus, popular research methods, the role of nondisabled participants and caregivers/specialists, and common research questions explored; Birhane et al. (2022a)’s survey of the state of the art in Machine Learning research within the top 100 most ‘influential’ papers (determined by highest citation count), finding that the research broadly values Performance, Generalization, Efficiency, Building on Past Work, and Novelty, with little to no discussion of potentially harmful outcomes or negative applications of the work; and Pang et al. (2025)’s exploration of LLM-related research within the CHI Conference on Human Factors, exploring the domains of HCI research where LLMs have been applied, the various roles performed by LLMs in such research and resultant types of contributions, and the limitations and risks of using LLMs as acknowledges by HCI researchers.

Such literature reviews are also common in NLP (e.g., Gao et al. 2023; Guo et al. 2024; Zhao et al. 2023). Our interests in exploring research around “bias” in LLMs/AI systems is perhaps closest to Bansal (2022)’s or Gupta et al. (2023)’s explorations of types/sources of [socio-demographic] biases within NLP research and mitigation strategies, Czarnowska, Vyas, and Shah (2021)’s work around fairness metrics used to measure “bias”, and Wan et al. (2024)’s study on gender, skin tone, and geocultural bias in T2Is, in terms of categorizing genres of bias within and taxonomizing harms caused by LLMs/AI systems.

Of particular note is the work of Blodgett et al. (2020), whose coverage of 146 papers studying “bias” in NLP systems revealed that an inconsistency around any conceptualizations of “bias” within the field. They provided a strong recommendation that researchers with similar goals should clearly define “bias” for their purposes, especially because the way in which they operationalize “bias” strongly influences how others will interpret their findings, i.e., researchers sharing interpretations of “bias” will be likely to follow each other. Their work, containing a checklist of questions to help define “bias”, has been strongly influential in the field, with over 1400 citations in less than 5 years (as of this writing), which is hopefully indicative of a good reception for future researchers similarly motivated to study “bias” in fields adjacent to NLP. One of the motivations of this paper is to translate and expand upon Blodgett et al. (2020)’s focus on NLP to research on LLMs/AI systems.

2.2 “Bias” in AI systems/LLMs

The Oxford English Dictionary defines “bias” as a ‘tendency to favor or dislike a person or thing, especially as a result of a preconceived opinion’ (Oxford English Dictionary 2025), or ‘to exert an influence on (a person or thing), often unduly or unfairly’ (Oxford English Dictionary Online 2025). It is immediately apparent from these definitions that to be biased or to have a bias carries a negative sentiment as operationalized towards a person or population, but that need not be the case. Indeed, the strictest sense of the word “bias” does not embed any sort of sentiment at all. It is a derivative of the Old French *biais* and Greek *epikarsios* (Online Etymology Dictionary 2025) simply meaning ‘slant’, often applied to cutting fabric along a diagonal. It entered the English language through the game of Crown Bowls to refer to the asymmetry of weight distributions of balls used in the game, affecting the curvature of the paths when rolled. This interpretation of “bias” can be seen in several of Shakespeare’s plays e.g. in *Troilus and Cressida* (Act IV, Scene 5, lines 7-8), he writes, “Blow, villain, till thy spherèd bias cheek Outswell the colic of puffed Aquilon,” where Ajax is being asked to blow his trumpet and puff out his cheeks, rounding them like biased balls in Crowns.

The idea of “bias” as a bad thing within statistics possibly began with the work of Ronald Fisher, who designed a statistical estimator to be consistent and measure the difference between true and estimated values of a parameter which he termed the degree of “bias” in reference to the idea of drifting away from a true result. Over time, this understanding of “bias” as a measurement of the difference between expected and true values grew to encode some form of intentional or unintentional attempts to create this difference, as concepts such as ‘researcher bias’ or ‘sampling bias’ were formulated. “Bias” thus became ‘a systematic distortion of an expected statistical result due to a factor not allowed for in its derivation’ (Olteanu et al. 2019), something to be eliminated.

Closer to modern-day fields of NLP and AI, Friedman and Nissenbaum (1996) were the first to explicitly address “bias” in computer systems as the tendency to ‘systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others.’ While they too added the negative valence to “bias”, they noted that for “bias” within designed systems to be discriminatory or harmful, it needed to be systematically applied along a pattern. They categorized such bias within computer systems as preexisting, technical, or emergent, and addressed the role of societal biases on the impacts of such systems including gender bias. At a time when research into the designing of large-scale NLP systems was ramping up, Friedman and Nissenbaum (1996) paved the way for modern-day AI ethics research by putting into focus how societal ‘biases’ affect computer systems.

Ascribing to the original concept of “bias” as simply a slant without any valence, it is more strictly accurate to compare it to having an opinion – everyone has their own opinions on a variety of topics, often informed by their lived experiences and positionalities (Gupta et al. 2023). Just as every individual has the right to hold their own opinions, the “idea of bias [within AI/LLM outputs] as something that can be eliminated, once and for all, is misleading and problem-

atic” (Birhane 2021). As Miceli, Posada, and Yang (2022) put it, “data never represents an absolute truth. Data, just like truth, is the product of subjective and asymmetrical social relations” (D’Ignazio and Klein 2020). However, the conception of “bias” simply as ‘prior knowledge’ of neutral valence remains uncommon within the field, with few exceptions (e.g., Bishop and Nasrabadi 2006; Caliskan, Bryson, and Narayanan 2017; Gupta et al. 2023).

On the other hand as “bias” is ‘harm’, which has no single definition due to its ubiquity across different fields. At a high level, ‘harm’ is a result of an activity ‘causing one or more of the following: pain, death, disability, mortality, loss of ability or freedom, and loss of pleasure’ (Gert 2004). Here, there is a much more clear sense of negative sentiment, where harm is believed to have occurred when an output produced by such tools or processes directly or indirectly results in bringing negatives consequences on an individual or a group. Closer to the AI/LLM context, harm can be *allocative* or *representational*. Proposed by Barocas et al. (2017), allocative harms occur where opportunities/resources are withheld from individuals/groups due to their identities, whereas representational harms follow unfairly-constructed depictions of individuals which may lead viewers to form negative stereotypes. Representational harms are categorized by Dev et al. (2020) into five types – *stereotyping*, or the overrepresentation of a set of beliefs about an identity, *disparagement*, or the idea that some groups of people are lesser than others, *dehumanization*, or the practice of treating certain groups of people as less than human, *erasure*, or the lack of representation of groups of people, and *quality of service*, or when models provide inequitable outcomes for different groups of people. Recent research into harms caused by the outputs of AI tools (e.g., Bianchi et al. 2023; Mack et al. 2024; Qadri et al. 2023) has mostly centered around representational harm.

Therefore, in the context of the datasets underneath and the outputs of AI systems and LLMs, **“bias” and “harm” are two different concepts that cannot be used interchangeably**. To be ‘biased’ is to simply lean towards or have a particular opinion, and the presence of such opinions within datasets or outputs of LLMs/AI systems cannot be eliminated. Because dataset biases are often reflections of the opinions of their creators and annotators, opinions typically driven by individuals with societal privilege, such biases “manifest existing power asymmetries” (Miceli, Posada, and Yang 2022), and cause harm upon those at the lower end of asymmetries. While the subtlety of this difference between “bias” and ‘harm’ might not be very consequential for the average user and public understanding of “biased LLMs/AI tools” is undoubtedly important towards creating effective change, we believe that researchers have a stronger responsibility to understand this distinction, especially in the pursuit of ‘debiasing’ endeavors towards some objective, fair, and unrealistic standard (Haraway 1988).

Thus motivated, we explore patterns of researchers studying “bias” within LLMs/AI systems, examining whether/how they define “bias” in their contexts and what aspects of identity they study “bias” against.

3 Methods

The methodology of this literature review is inspired by similar research such as Ali et al. (2023), Birhane et al. (2022a), Gupta et al. (2023), Mack et al. (2021), and others. ‘The field’ is first narrowed down to specific conferences and publication societies where research around “bias” in AI systems is commonly published, based on our experience as authors publishing in this field. The largest such venue is the Association of Computational Linguistics (ACL), which is where the search began: by downloading the metadata for the entire ACL anthology of 105,862 papers. This is an appropriate way to search the ACL anthology instead of using the native Search bar, which does not allow for robust Boolean search and targeted downloads like other databases. After this collection, all papers published before 2014 are excluded, since the past 10 years (as of this writing) is a sufficient range of time within which to study the development of AI systems, narrowing down to 67319.

This set of papers was augmented by including papers from three other conferences/venues identified as relevant: the ACM Conference on Fairness, Accountability, and Transparency (FAccT), the proceedings of the Advances in Neural Information Processing Systems (NeurIPS), and the Association for the Advancement of Artificial Intelligence (AAAI). FAccT proceedings are indexed on the ACM Digital Library and a search revealed 734 entries, all of which have been published after 2019 because the first ever FAccT conference (then referred to as FAT*) was held in 2018 and proceedings from the 2019 conference onwards are indexed on the ACM Digital Library. Searching for NeurIPS (once called NIPS) and AAAI papers was harder because their respective databases are less robust than the ACM Digital Library, so those proceedings were searched for on Google Scholar. All proceedings for both conferences were retrieved with Google Scholar keywords ‘site:proceedings.neurips.cc’ and ‘site:ojs.aaai.org’ resulting in 26400 and 25200 results, respectively. Applying the filter of papers being published in 2014 or later, both numbers came down to 21100.

This set of papers was subjected to a keyword search, designed to streamline the corpus towards the stated field. The keyword search was applied directly to paper titles and Abstracts for ACL papers, and used as a web search query for the other venues. The query was as follows:

```
'ai' OR 'artificial intelligence'  
OR 'llm' OR 'large language model'  
OR 'AI') AND ('bias' OR 'harm' OR  
              'stereotype')
```

This resulted in 208 matches from the ACL Anthology, 545 matches from FAccT, 3980 from NeurIPS, and 3350 from AAAI. To simplify the screening process for Google Scholar results from NeurIPS and AAAI proceedings, the search only considers the first 25 pages (250 results) for each of these results. Furthermore, the full set of 545 results from FAccT are downloaded and the query is re-applied to Titles and Abstracts (since ACM Digital Library performs a ‘deep search’ into contents of documents, and not just restricted to Titles and Abstracts), reducing the set to 19 papers. In total,

the screening dataset contains 727 papers from four conferences for abstract screening. These papers were manually screened to see if they answered the following questions:

- Does the work study existing AI/LLM systems, or processes within them?
- Does this paper study human identity-biases within or harms caused by AI technologies or LLMs?
- Does the work produce new (fine-tuned) AI systems/LLMs, or ways to improve ones?

We focus on research on “bias” in contexts of aspects of human identity or the presence of any sort of slant/opinion within content (e.g., political bias), which can then cause harm in downstream services, as opposed to other types such as position bias, frequency bias, recency bias, demonstration bias and other types of biases that are not based on aspects of human identity (e.g., Diehl Martinez et al. 2024; Li et al. 2024; Stacey et al. 2020). Papers studying such biases are screened out, alongside others where LLMs were used to form or augment datasets for some purpose such as debiasing (e.g., Han et al. 2024), research around specific reasons why biases might exist such as the impact of token lengths or (e.g., Ueda et al. 2024), and work around building fairer ML systems without focusing on how such systems are unfair (e.g., Zeng et al. 2024). Papers were included if they discussed the presence of bias within and harms caused by the outputs of AI tools and LLMs in a variety of ways, such as bias detection, debiasing and finetuning within models or AI tools. Papers were marked as Yes or No for this question, erring on the side of caution by including inconclusive papers at this stage and allowing for their exclusion during a second round of review.

Other examples of papers excluded are research on using LLMs to write medical systematic reviews (Ji et al. 2025; Yun et al. 2023), using LLMs to analyze memes (Sharma et al. 2023) or fairy tales (Toro Isaza et al. 2023), literature reviews (Bertrand et al. 2022; Blodgett et al. 2020) or reviews of service providers (Raghavan et al. 2020), and tutorials (Chang, Prabhakaran, and Ordóñez 2019), to name a few. Two rounds of abstract review resulted in a final set of 189 papers selected for full review. A visual representation of the search and screening process can be seen in Figure 1, and the full set of papers is shown in the full version of the paper on arXiv. Data collection and annotation was conducted in January 2025.

The 189 selected papers were then analyzed through the following series of questions, towards the stated goal of understanding how the researchers addressed bias and harm within AI systems and LLMs:

1. What kind of bias(es) is/are studied?
2. Is “bias” defined? If so, how?
3. Does the paper address ‘harm’? If so, is “bias” differentiated from ‘harm’?
4. Does the work propose design recommendations for future AI/LLM systems, or otherwise inform usage of such systems by regular users?

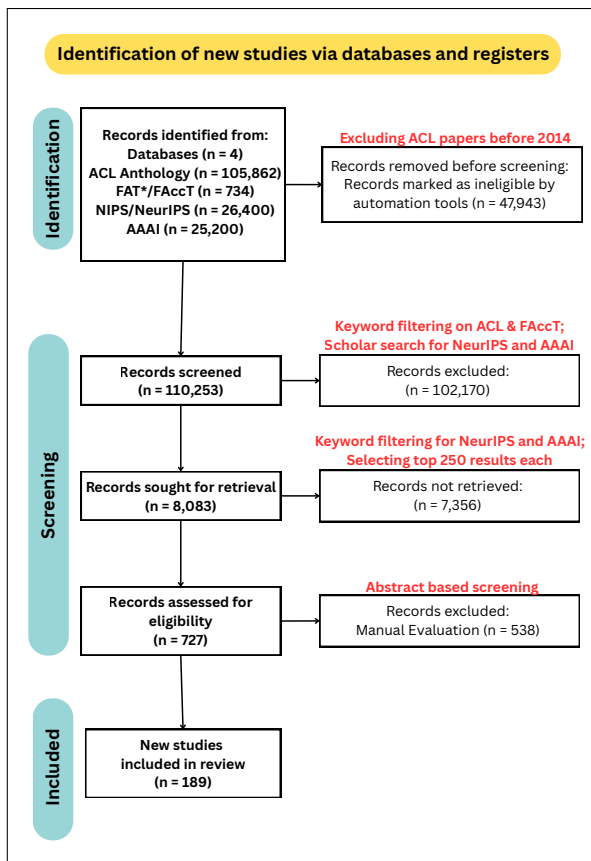


Figure 1: PRISMA diagram of sampling and filtration.

4 Findings

4.1 “Bias” as an Undefined Negative Aspect

One of the most prominent findings is that 82% (155/189) of surveyed papers did not provide a definition of “bias”, or related terms such as ‘fairness’, as they had conceived it. These authors often acknowledged the presence of bias within LLMs/AI systems and pointed to previous work highlighting such biases, before moving into their own work.

For 61/155 papers, we observed that their authors would provide examples of what it meant for models to be biased, instead of definitions. A few excerpts are shown below:

“A sentence that highlights gender bias is: The doctor told the nurse that she had been busy. A human translator carrying out coreference resolution would infer that ‘she’ refers to the doctor... [but] an NMT model trained on a biased dataset in which most doctors are male might incorrectly default to the masculine form” - (Saunders and Byrne 2020)

“...subtle biases that a language model may harbor (e.g., a tendency to talk more about clothing and cooking with regard to women)” - (Dwivedi-Yu 2024)

“...if the sentence distributes its attention on social groups differently (e.g. ‘doctor’ attends to ‘he’ and not to ‘she’), then there is bias.” - (Gaci et al. 2022)

Within the remaining 94/155 papers, “bias” is either evaluated or defined purely mathematically, where researchers establish a metric for calculating “bias” and then mention how a score on that metric is indicative of “bias” e.g.:

“A variable Y in a causal graph exhibits potential proxy bias if there exists a directed path from D to Y that is blocked by a proxy variable P and if Y itself is not a proxy.” - (Ding et al. 2022)

“We define the gender bias in the retrieved image set is quantified as the normalized absolute difference in counts of each gender’s images.” - (Ghate, Choudhry, and Bannihatti Kumar 2024)

“We introduced the Normalized Stereotype Score (nss) defined as follows: $nss = \frac{\min(ss, 100 - ss)}{0.5}$. Hence, nss is a measure that stays between 0 and 100 where 100 is the non-biased value.” - (Ranaldi et al. 2024)

On the other hand, only 18% (34/189) papers provided explicit definitions of “bias” for their use cases, without referencing mathematical definitions. Though some leaned on previous definitions provided by work in the field (e.g., by Baker and Hawn 2022; Dev et al. 2021; Sap et al. 2020, to name a few), we highlight a few examples where researchers provided their own clear definitions of “bias”:

“Bias in AI, in an ethically significant sense, refers to the tendency of AI actions (e.g., decisions, generations) to favour one individual or group over others in ways that deviate from accepted standards.” - (Guan et al. 2025)

“We primarily categorize biases into semantic-related and semantic-agnostic biases... Semantic-related bias pertains to the bias of evaluators that is affected by elements related to the content of the text... Semantic-agnostic bias refers to the bias of evaluators that is influenced by factors unrelated to the semantic content of the text.” - (Chen et al. 2024)

Irrespective of whether bias is defined, another emergent pattern is that while it is generally apparent that “bias” is something negative to be removed (evidenced by the performance of ‘debiasing’), this negative connotation is explicitly made clear in 32 cases. Some excerpts are below:

“Social bias can be defined as the manifestation through language of prejudices, stereotypes, and discriminatory attitudes against certain groups of people.” - (Marchiori Manerba et al. 2024)

“Bias in NLP applications makes distinct judgments on people based on their gender, race, religion, region, or other social groups could be harmful, such as automatically downgrading the resumes of female applicants in recruiting.” - (Li et al. 2022)

“We define gender bias as undesirable variations in how the model associates an entity with different genders.” - (Srinivasan and Bisk 2022)

By attaching a negative sentiment to it, several papers conflate “bias” with ‘harm’. We highlight a few exceptions

that acknowledge how bias and stereotypes do not have a valence by default, and ‘debiasing’ isn’t always optimal:

“Bias is not inherently always bad, and to say that T2Is exhibit bias is not a bad thing. Humans have biases, and therefore it is but natural that human-designed systems embed biases, and in turn apply them in their operations. Such biases are not inherently negative ... and it is only when they embed unjust stereotypes about people or groups that they can cause harm.” - (Ghosh, Lutz, and Caliskan 2024)

“Stereotypes shape social perceptions — either positive or negative prejudices and pre-existing judgments about particular groups and the people who belong to them without any objective basis” - (Shin et al. 2024)

“Stereotypes are a positive or negative, generalized, and often widely shared belief about the attributes of certain groups of people” - (Herold, Waller, and Kushalnagar 2022)

“Stereotypes refer to common generalizations about the qualities of people based on their associations with groups and whether they are positive or negative could have different implications.” - (Mei, Fereidooni, and Caliskan 2023)

“One should also be careful in the use of debiasing. Removing signals about race or gender ... may also remove key features of models needed for analyses. For example, removing gender or race ‘signal’ from the model may severely hamper the use of that model in gender studies or work on critical race theory.” - (Yifei, Ungar, and Sedoc 2023)

Finally, while Blodgett et al. (2020) is influential within this corpus for a wide range of reasons – mostly the definitions of allocational and representational harms – we identify 3 papers (post (Blodgett et al. 2020)’s publication) where researchers provide a definition of “bias” explicitly referencing Blodgett et al. (2020)’s directive.

“We must first define the theoretical construct of social bias. In contrast, Blodgett et al. (2020) showed many works in NLP failed to (adequately) define social bias.” - (Bommasani and Liang 2024)

“Within this context, specifying which biases to analyze is crucial; Blodgett et al. (2020) find that a majority of NLP papers investigating bias are unclear in their articulations of bias. In this paper, we consider both representational and allocational harms” - (Kirk et al. 2021)

“Following the recommendations of Blodgett et al. (2020), we explicitly define gender bias as the tendency of these models to generate or perpetuate gender stereotypes.” - (Ffririksdóttir and Einarsson 2024)

4.2 Overrepresentation of Gender Bias Research

Within our corpus of papers, we observe that 79.9% (151/189) papers focused on ‘gender bias’. 30.7% (58/189) papers focused exclusively on gender bias and out of these,

37 (19.6% of the corpus) focused on gender and occupation biases i.e., exploring patterns of LLMs/AI tools associating specific occupations with genders. In 18% (34/189) papers, researchers studied the intersection of gender and race/ethnicity (e.g., Arzaghi, Carichon, and Farnadi 2024; Guo and Caliskan 2021; Steed and Caliskan 2021; Tan and Celis 2019), while 31.2% (59/189) papers explored intersections of gender with other aspects of identity.

Some researchers provide rationale around why they focus on gender bias, with the most common being that gender bias is one of the easiest types of bias to study, given the existence of large datasets with labeled gender data. Therefore, researchers often use gender bias as an example or proxy for other biases, arguing that their findings along reducing gender bias can apply to other social biases. In their own words:

“Although this paper focused on gender bias, it is relevant to work examining other forms of bias, such as racial stereotyping, in embeddings.”
- (Zhang, Sneyd, and Stevenson 2020)

“We consider gender bias as a running example throughout this paper and evaluate the proposed method with respect to its ability to overcome gender bias in contextualised word embeddings, and defer extensions to other types of biases to future work.”
- (Kaneko et al. 2022)

Within the 79.9% (151/189) papers that focus on gender, 138 consider exclusively binary constructions of gender and 114 of these mention such a binary focus as a limitation of the work. This is often explained to be the case because labeled datasets, popularly used for ease of research, often only contain data on binary gender. As examples, consider:

“Due to practical reasons and existing lack of datasets, we limited our research to only the binary genders.” - (Jeyaraj and Delany 2024)

“First, due to the limited amount of datasets and previous literature on minority groups and additional backgrounds, our study was only able to consider the binary gender when analyzing biases.”
- (Wan et al. 2023)

“We investigated under the binary gender setting, because of the limitation of the existing benchmarks.”
- (Luo and Glass 2023)

Examples of papers that explore constructions of gender beyond the binary include Cabello et al. (2023), Ghosh and Caliskan (2023), Hall et al. (2023), and You et al. (2024), to name a few. We also note that Dennler et al. (2023) focuses on harmful effects of AI systems against queer users while also incorporating conversations of gender within their work, though the two are not the same.

The overrepresentation of gender bias in LLM/AI bias research is also not lost on researchers within our corpus, as can be seen from a few representative examples:

“Recent work on language models show substantial evidence of the presence of sociodemographic biases associated with race and gender ... However, little

prior work has focused on the identification and impact of disability bias.”

- (Venkit, Srinath, and Wilson 2022)

“For instance, much work focuses only on mitigating gender bias despite pre-trained language models being plagued by other social biases.”

- (Meade, Poole-Dayana, and Reddy 2022)

“Existing works on biases in machine translation have almost exclusively focused on issues of gender biases.” - (Huang and Xiong 2024)

Apart from gender bias, the most-focused areas were race/ethnicity bias (30.2%; 57/189 papers), nationality bias (13.2%; 25/189 papers), age bias (20.6%; 39/189 papers), and religion bias (19.1%; 36/189 papers).

It is thus evident that the research captured within this corpus narrowly focuses on gender and a few other aspects of human identity as they examine ‘bias towards marginalized populations,’ overlooking several others. This is also noted by a few researchers within our corpus:

“Several studies have investigated LLM bias and harm, they predominantly focused on racial and gender biases in language models— dimensions that dominate Western public discourse. Few works have explored harms and stereotypes in the Global South contexts.” - (Dammu et al. 2024)

“Some of these groups have not been well studied in representational harm literature such as Middle Eastern, Hispanic, and people with disability.”

- (Hosseini, Palangi, and Awadallah 2023)

There are, however, a few papers that focus on bias relevant to identities beyond the aforementioned: Dammu et al. (2024) and Ghosh (2024) explored bias in the context of caste within India/the Indian diaspora; Mujtaba et al. (2024) studied bias against speakers who stutter within automatic speech recognition systems and the LLMs that drive them; Abboud and Oz (2024) examined how LLMs exhibit bias against lesser-known dialects of languages such as Arabic and German; and Wolfe et al. (2024) focused on bias against adolescents and teenagers within AI systems, to name a few.

4.3 Lack of Implementation Plans within Research

Finally, perhaps the strongest finding across our corpus is that 89.4% (169/189) papers do not include ways to operationalize their proposed mitigation strategies or recommendations of other strategies such that the biases they studied are not propagated in real-world AI systems. Notably, we do not conflate the absence of implementable debiasing methods with the lack of production of novel methods: indeed 86.2% (163/189) papers do design some form of novel debiasing method or novel benchmark that can be used to train models to not produce biased outputs. In most of such cases, we also note that researchers often publish their code and training datasets on the Internet for public usage. However, we argue that there is a difference between

this practice of releasing code/datasets and providing actionable steps for implementing one’s research: one is for academic transparency, and the other is design responsibility, which involves clearly laying out the specific contexts and conditions under which the provided method can be used (Trisovic et al. 2022). cursory glances at linked codebases within papers revealed them to either provide explanation of the contents of various files within code and how to execute them (e.g., Amrhein et al. 2023; Han et al. 2022; He et al. 2022), or in rare cases, not much additional information at all (e.g., Adewumi et al. 2023; Jha et al. 2024).

A strong outlier is Wu, Nakashima, and Garcia (2024), who demonstrated gender bias in Stable Diffusion outputs beyond representations of human faces and the correlation between nouns in prompts to gender biases. They included a set of recommendations for Stable Diffusion and other text-to-image generator developers to reduce gender bias, and for users to practice specific prompting strategies based on their work to avoid received gender-biased outputs from Stable Diffusion until model iterations improved. Another example is how Luccioni et al. (2024) uncovered several patterns of biased outputs within popular text-to-image generators Stable Diffusion and DALL-E along with comparisons of their performance in response to similar prompts, while also developing public-facing tools like Diffusion Bias Explorer (available through HuggingFace) for users to see how these AI tools depicted different professions and adjectives, such that they can make informed decisions about a choice of tool for their own purpose. Leiding and Rogers (2024) supplemented their findings around how state-of-the-art LLMs were practicing safety behavior around stereotyping harms with a conversation on how policymakers could prioritize social impact as a measurement metric for LLM leaderboards, thus informing users which LLMs are ‘safer’. Finally, Kay, Kasirzadeh, and Mohamed (2024) introduced the theory of ‘generative algorithmic epistemic injustice’ by analyzing how the outputs of generative AI tools caused epistemic harm upon marginalized populations, and outlined a path towards Generative Epistemic Justice through actions for future developers of AI tools.

5 Discussion

5.1 The Dangers of “Bias” == ‘Bad’

One of the most prominent themes across our findings was that 82% (155/189) papers studying “bias” in LLMs/AI systems did not explicitly define what “bias” meant for their contexts, and conceive of it as a negative attribute to be removed. This is made clear through a large section of papers pursuing ‘debiasing’ research (thus treating “bias” as something negative to be removed) or explicitly making it known by using words such as ‘undesirable’ (Srinivasan and Bisk 2022) or ‘discriminatory’ (Marchiori Manerba et al. 2024).

It is important to tease apart the nuance of why “bias” == ‘bad’ is an issue. While it is undoubtedly true that many patterns (such as gender-occupations ones of man == doctor, woman == nurse) within the outputs of AI systems/LLMs are problematic and in need of mitigation, it would be untrue to make a blanket association of “bias” as being *always* neg-

ative, because some biases can be for social good. Consider a hypothetical model that studies demographic data from a group of people and, in deciding who should get flu shots first, picks older adults ahead of younger adults and children ahead of adolescents. While such a model can be labeled as practicing age bias, this outcome likely factors in the higher probability of older adults and young children being immunocompromised and/or having a higher risk of contracting the flu, i.e., an age-blind approach treating people equally would be unfair upon older adults and young children. Therefore, researchers should carefully consider that the “bias” being investigated does/can create *inequitable* (not unequal) outcomes, before labeling it as ‘bad’.

A simplistic idea of “bias” == bad also sets designers down a road of debiasing that has an unachievable end. As mentioned before, to have “bias” is to simply have an opinion, and the act of debiasing seeks to eliminate opinion-based outcomes towards some paradigm of objectivity, thus fueling the argument that objective AI systems must replace biased humans in decision making capacities to attain true equality. In reality, such an objective model is impossible: humans leave crumbs of their biases within individual choices at various stages of the model design process (Suresh and Guttag 2021), and such individual choices interact in complex ways to shape the outputs of LLMs/AI systems, and these choices are inextricably linked with the models themselves. While researchers can design innovative approaches to reduce “bias” along one/few aspects of identity, it is impossible to remove all “bias” from LLMs.

Further, debiasing by removing attributes such as race and gender amount to ‘whitewashing’ models, thus having a negative impact on use of such a model for critical work (Yifei, Ungar, and Sedoc 2023). Consider the finding that the 2013 LA County Coordinated Entry System assessing homeless individuals’ vulnerability through an algorithm ended up discriminating along racial lines by systematically considering White adults as more vulnerable and in higher need of permanent housing than Black or Latino adults (Varner and Lecher 2023). Removing race signals from datasets entirely might suggest that decisions cannot be made on the basis of race, but also does not account for the fact that application packets and the lives of applicants will undoubtedly have been shaped due to the opportunities they have been previously afforded, opportunities that historically have been affected by racial identity. This is equivalent to the social theory of colorblindness – “the belief that racial group membership should not be taken into account, or even noticed” (Apfelbaum, Norton, and Sommers 2012) – situated in the larger context that if aspects of identity are ignored in conversations, there will no longer be discrimination along those lines. This is a dangerous line of thinking towards fairness, striving to achieve mathematically equal outcomes for all groups (in this case, people of all genders), because such a movement is not equitable and does not consider the societal structures and events that led to dataset imbalances and sources of “bias” that then emerge as outputs (Birhane et al. 2022b). While the lay user need not concern themselves with the intricacies of the terminology but as researchers, we owe it to ourselves to not oversimplify “bias” as *always* bad.

5.2 “Bias” == Gender Bias, and ‘Fairness Gerrymandering’

Another observed trend was that 79.9% (151/189) papers that studied “bias” within LLMs/AI systems explored gender bias, such that it is overrepresented in the corpus relative to studies focusing on bias based on other aspects of identity. Researchers mention doing so because the availability of several gender-labeled, specifically binary gender, datasets made it easy and feasible to study (e.g., Jeyaraj and Delany 2024; Wan et al. 2023). This focus on gender bias was also recognized as disproportionately high by researchers within our studied sample, such as Venkit, Srinath, and Wilson (2023) and Meade, Poole-Dayana, and Reddy (2022).

A disproportionately strong focus on gender bias can create conditions whereby the field of AI bias and the volume of research *actively causes harm* upon sections of historically marginalized populations: by focusing with disproportionate strength on one aspect of bias (binary gender), researchers tend to design for the uplifting of small sections of marginalized populations (in this case, individuals identifying as female), leaving others behind (Ghosh et al. 2025). The gulf between communities, in terms of representation, thus widens and creates situations where sections of historically marginalized populations (such as caste-oppressed communities in India, speakers who stutter, speakers of low-resource languages and lesser-known dialects) do not see their representation or equitable treatment in models at all improved. This leads to the development of LLMs/AI systems that practice intra-marginal inequality, whereby they are ‘fair’ based on one or a few axes of identity, namely binary gender, but remain biased (possibly more than popular baselines) through what Kearns et al. (2018) calls ‘Fairness Gerrymandering’. With the surging development of novel LLMs and AI systems built upon their backs for global usage, continuing a narrow focus on a select few aspects of human identity in bias research is a recipe for disaster. There are strong parallels to this in society, as US Affirmative Action laws, while intended to address historical discrimination and promote diversity in education and employment, disproportionately benefits white women (Kohn 2013) and the system has mostly focused on mitigating gender bias.

Studying gender bias ‘as an example’ (e.g., Zhao et al. 2017) of bias within LLMs/AI systems to study and develop debiasing techniques around also represents a problematic situation. Such approaches sometimes boil ‘gender’ down to a property that can have (often) two mutually-exclusive labels where one of those labels is often associated with some other trend within another category (e.g., professions), which ignores the complex social constructions of gender and the real-world inequities (such as hiring gaps) that have led to such associations being formed. This danger can extend if researchers consider debiasing approaches effective for reducing gender bias to be able to extend to other binary identities, such as (as often conceptualized) disabled/non-disabled or, in a much more alarming sense, White/Black.

5.3 The ‘Last Mile’ Gap: Research Findings Don’t Translate into Real-World Systems

We also notice within our corpus that while 89.4% (169/189) researchers explored how AI systems/LLMs are biased and/or developed demonstrably successful debiasing techniques, they did not include any information about potential steps for designers of real-world AI systems built on LLMs they studied (such as models within the BERT family, the GPT family, and others available through HuggingFace) to implement their findings so that systems would exhibit little to no bias. This deepens ‘the last mile gap’ (Cabitza, Campagner, and Balsano 2020), where debiasing advances from academic research are rarely implemented in real-world systems, which continue to exhibit ‘bias’.

This is non-trivial: researchers study mathematical debiasing methods in controlled conditions and it cannot be assumed that designers of real-world AI systems built on the LLMs being studied will find replication of research obvious. Additionally, researchers developing ‘debiasing’ methods within our corpus often focus on individual associations, such as gender-occupation associations, and sometimes make simplistic assumptions on how their work can improve the outcomes in complex downstream tasks. This is also true of model usage, especially for research that considers debiasing in single-model contexts and does not consider how real-world AI systems might be implemented in multi-agent architectures with complex interactions of models.

We acknowledge that a likely cause behind a lot of researchers not including any mention of potential implementation strategies for their findings could be the fact that the venues under consideration – *ACL, FAT*/FAccT, NIPS/NeurIPS, and AAI – have publication rules around the maximum length of papers. Researchers likely prioritized the inclusion of important figures/tables conveying information about their work and results, and did not have a lot of permitted space for additional information about implementation. Though we are sympathetic to such a plight having made decisions around the best use of real estate on the page when working under the page constraints of these venues, we still argue that at least some content about potential implementation plans around the findings the researchers so painstakingly detail should have been prioritized.

We must also add two caveats here: firstly, we cannot be certain that none of the findings and steps undertaken to debias LLMs/AI systems present within our corpus made it into any LLMs/AI systems currently available for public/private usage, because research papers likely will not document conversations between researchers and developers after the publication of such findings. We simply note the widespread absence within written papers of actionable steps proposed by researchers to implement their findings into real-world systems, as we have argued for why they are important to include. It is also unreasonable for researchers to preempt *all* questions designers of real-world AI systems might have towards implementing their work, and therefore implementation plans might be incomplete in several ways. That said, we believe that an incomplete implementation plan is better than none at all, to get developers interested.

6 Recommendations for Future Research

6.1 Define and Expand Focus of “Bias” Coverage

We encourage researchers within the field to expand their coverage in terms of populations against whom the outputs of LLMs/AI systems are ‘biased’. While we noted the overrepresentation of gender bias within our corpus and sparser coverage of other aspects of identity (such as race, nationality, religion, and age), there remain several additional aspects of identity that receive very little focus in this vein. In particular, we note Qadri et al. (2023) and Ghosh, Lutz, and Caliskan (2024)’s observation that studies of AI/LLM bias often exclude aspects of identity prevalent within non-Western contexts or lenses of analyzing globally-held identities through non-Western lenses. We encourage further research around such populations at ‘the forgotten margins’ (Birhane et al. 2022b) of AI/LLM fairness research.

The unavailability of large datasets, often the reason for keeping studies around gender bias restricted to the binary, should not be considered a barrier in terms of studying bias. For instance, Ghosh (2024) studied casteist bias within the text-to-image generator Stable Diffusion through its outputs, and still demonstrated patterns at scale. Especially in an age where large-scale generative AI systems are being used by millions of users across the world, the reliance on existing datasets and benchmarks to study bias should be reduced.

Furthermore, we invite researchers to expand studies of “bias” into ‘harm’ by consulting users who encounter AI/LLM outputs “biased” along aspects of their identity. While also solving the problem of lacking data by offering empirical lived experiences, users are a much better indicator of the *impact* of such “bias”. While a majority of researchers make [paraphrased] statements such as ‘bias in AI can have devastating outcomes on marginalized people’ and operate under the assumption that because they have a passable understanding of such outcomes, they can simply focus on technical solutions and demonstrating “bias” at scale, the fact remains that because AI systems are reaching new populations at a hitherto-unseen rate, that is no longer a fair assumption. We advocate for further investigations of “bias” within AI/LLM outputs along aspects of identity that are currently vastly overlooked in such research. We encourage conferences such as AIES and FAccT to invite thus-motivated work through targeted calls and special tracks, while also creating necessary conditions that make participation (often held in Western countries and in English, serving language and cost-based barriers) equitable.

6.2 (Partially) Discuss Implementation Strategies

Recognizing the academia-industry gap in terms of researchers not discussing actionable implementation plans for their findings into real-world AI systems, we align ourselves with Gupta et al. (2023) and Venkit et al. (2023) as we advocate for stronger discussion of such implementation steps, even if they are imperfect and not all-encompassing.

Researchers can create some fact sheets for this purpose. Mitchell et al. (2019) proposed this idea for publishing additional information alongside novel ML models, describing intended use, relevant factors, performance metrics used,

Fact Sheet

High-level Facts

- What aspect(s) of bias is/are mitigated?
- What dataset(s) and model(s) is/are used in experiments?
- How is “success” defined and achieved?
- What methods (qualitative, quantitative, mixed) are used?
- (If human subjects), describe positionality of subjects/annotators.

Limitations and Ethical Considerations

- What are the stated limitations of this research?
- What potential risks can occur if this work is implemented?
- What subpopulation(s) of users is/are the most likely to experience risks?
- How might potential risks cause harm upon affected users?
- What are some mitigation strategies against the potential risks?
- How will strategies differ across affected subpopulations?
- Has this method been tested for other potential amplification of bias(es)?

Intended Use Cases

- In what real-world systems can this work be implemented?
- What real-world AI systems is this work inappropriate for?
- Is the work applicable for a definition of bias within specific cultures and contexts, or can it be applied in global systems?
- In what languages is this tested, and can it be extended?

Implementation-level Details

- What computational resources were required to execute this experiment (e.g., GPU hours, memory, latency tradeoffs, etc.)?
- Have experiments addressed performance in multi-agent settings? How did the implementation of this method affect performance?
- Does the method perform debiasing at training- or inference-time?
- Does the implementation of this method require any retraining baseline model(s) from scratch or modifying any internals?
- Are available experiment details and code clearly documented for independent replication of results?

Maintenance and Lifecycle

- Are there likely scenarios where this method might break down?
- Are there maintenance tasks that are required for the long-term success of the implementation of this debiasing method?
- Are there components that can be expected to deprecate over time?
- Does the system support human-in-the-loop extensions post-deployment?

Contact <researcher name> at <email> for more information!

Figure 2: Suggested prompts for designing a Fact Sheet.

datasets leveraged for training/testing, ethical implications, and caveats, among other things. Along the same lines, researchers can provide a brief one-page summary of their research and brainstorm some implementation-level questions that potential AI designers might have. We provide a sample fact sheet in Figure 2. This Fact Sheet allows researchers to lay out a few basic details about their work, as well as consider a few implementation-level questions such as Intended Use (e.g., chatbots, decision making systems, language learning tools, etc.), Limitations and Ethical Considerations (e.g., potential risks, affected populations, and mitigation plans), Implementation details (e.g., potential computational cost comparisons) and Maintenance information (e.g., deprecation of resources). We also acknowledge that this set of questions might be incomplete, and invite researchers to provide additional information as appropriate.

6.3 Stronger Collaboration between Academic Research and Industry

To mitigate the ‘last mile’ gap in reducing “bias” in AI systems/LLMs and the continued presence of such “bias” in real-world AI systems/LLMs being developed in industry, we encourage a stronger collaboration between the two.

We believe it unfair to ask one group or the other to entirely shoulder the burden of producing “debiased” AI systems/LLMs: academic researchers rarely have the capital or workforce required to design and deploy a large-scale system, and industry practitioners often are unable to identify the ways in which their systems can contain “bias” during development (Raji et al. 2020). While the role of academic researchers as investigators into complex sociotechnical challenges within designed AI systems/LLMs is undoubtedly important for unearthing to the common user the ways in which systems can be harmful, we believe that such academic researchers can and do also have a role to play in the design of ‘better’ and ‘safer’ systems. We thus encourage active synergistic efforts between academic researchers and industry practitioners who design AI systems/LLMs, towards the development of AI systems that mitigate much of the “bias” highlighted by this corpus of papers.

7 Limitations and Future Work

One limitation is that the corpus might have overlooked relevant work. For instance, this search did not cover CHI, a reputable conference but not one for publishing research around AI/LLM bias/harm, since less than 4% papers over the past 5 years has been LLM-related (Pang et al. 2025)). Thus, papers such as Mack et al. (2024) might have been missed.

Further, while screening papers based on Titles/Abstracts is consistent with prior literature reviews (e.g., Ali et al. 2023), it relies on authors *using exactly those words*. This is another avenue through which papers could have been overlooked, as authors could have used other words to refer to the same concepts. As an example, Qadri et al. (2023)’s paper titled *AI’s regimes of representation: A community-centered study of text-to-image models in South Asia*, did not mention ‘bias’, ‘harm’ or ‘stereotype’ in the title and was thus missed. We experimented with changing the AND in the Boolean filter to OR, and a random sample of 50 papers from this alternate query revealed that 71% would have been manually screened out, so we did not amend the query.

We also recognize the trend within the field to upload research on archival sites such as arXiv, as researchers prioritize faster visibility of research, and societal power and privilege structures (e.g., conferences such as this one requiring in-person participation for publication cater towards researchers with the financial means to afford travel/stay, over and above visa costs for international applicants) prevent several top researchers from submitting their work for peer-review at the conferences studied here. Nevertheless, the coverage quite possibly excludes a section of work within the field (495 papers on arXiv matched our query, and 12% of these papers are already captured in this review) which might have affected our findings.

This literature review also opens up some avenues for future work. While this search screened for research around ‘bias’ or ‘harm’, it could also be interesting to examine how “debiasing” research occurs in the AI ‘safety’ field, and whether work around making “safer” AI systems is different in their coverage of marginalized populations and/or less-prone to a ‘last mile’ gap. Additionally, similar reviews could cover relevant journals e.g., AI & Society.

Acknowledgments

The findings of this paper do not necessarily reflect views of all of the authors. We are grateful to our reviewers for their thoughtful feedback which undoubtedly made this paper stronger!

References

- Abboud, K.; and Oz, G. 2024. Towards equitable natural language understanding systems for dialectal cohorts: Debiasing training data. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 16487–16499.
- Adewumi, T.; Södergren, I.; Alkhaled, L.; Al-azzawi, S.; Simistira Liwicki, F.; and Liwicki, M. 2023. Bipol: Multi-Axes Evaluation of Bias with Explainability in Benchmark Datasets. In Mitkov, R.; and Angelova, G., eds., *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, 1–10. Varna, Bulgaria: IN-COMA Ltd., Shoumen, Bulgaria.
- Ali, M.; Ghosh, S.; Rao, P.; Dhegaskar, R.; Jawort, S.; Medler, A.; Shi, M.; and Dasgupta, S. 2023. Taking Stock of Concept Inventories in Computing Education: A Systematic Literature Review. In *Proceedings of the 2023 ACM Conference on International Computing Education Research*, volume 1, 397–415.
- Amrhein, C.; Schottmann, F.; Sennrich, R.; and Läubli, S. 2023. Exploiting Biased Models to De-bias Text: A Gender-Fair Rewriting Model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4486–4506. Toronto, Canada: Association for Computational Linguistics.
- Apfelbaum, E. P.; Norton, M. I.; and Sommers, S. R. 2012. Racial color blindness: Emergence, practice, and implications. *Current directions in psychological science*, 21(3): 205–209.
- Arzaghi, M.; Carichon, F.; and Farnadi, G. 2024. Understanding Intrinsic Socioeconomic Biases in Large Language Models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 49–60.
- Baker, R. S.; and Hawn, A. 2022. Algorithmic bias in education. *International journal of artificial intelligence in education*, 1–41.
- Bansal, R. 2022. A survey on bias and fairness in natural language processing. *arXiv preprint arXiv:2204.09591*.
- Barocas, S.; Crawford, K.; Shapiro, A.; and Wallach, H. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual conference of the special interest group for computing, information and society*.
- Bertrand, A.; Belloum, R.; Eagan, J. R.; and Maxwell, W. 2022. How cognitive biases affect XAI-assisted decision-making: A systematic review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 78–91.
- Bianchi, F.; Kalluri, P.; Durmus, E.; Ladhak, F.; Cheng, M.; Nozza, D.; Hashimoto, T.; Jurafsky, D.; Zou, J.; and Caliskan, A. 2023. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1493–1504.
- Birhane, A. 2021. Algorithmic injustice: a relational ethics approach. *Patterns*, 2(2).
- Birhane, A.; Kalluri, P.; Card, D.; Agnew, W.; Dotan, R.; and Bao, M. 2022a. The values encoded in machine learning research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 173–184.
- Birhane, A.; Ruane, E.; Laurent, T.; S. Brown, M.; Flowers, J.; Ventresque, A.; and L. Dancy, C. 2022b. The forgotten margins of AI ethics. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 948–958.
- Bishop, C. M.; and Nasrabadi, N. M. 2006. *Pattern recognition and machine learning*, volume 4. Springer.
- Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (Technology) is Power: A Critical Survey of “Bias” in NLP. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476. Online: Association for Computational Linguistics.
- Bommasani, R.; and Liang, P. 2024. Trustworthy social bias measurement. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 210–224.
- Cabello, L.; Bugliarello, E.; Brandl, S.; and Elliott, D. 2023. Evaluating Bias and Fairness in Gender-Neutral Pretrained Vision-and-Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 8465–8483. Singapore: Association for Computational Linguistics.
- Cabitz, F.; Campagner, A.; and Balsano, C. 2020. Bridging the “last mile” gap between AI implementation and operation: “data awareness” that matters. *Annals of translational medicine*, 8(7): 501.
- Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186.
- Chang, K.-W.; Prabhakaran, V.; and Ordonez, V. 2019. Bias and fairness in natural language processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*.
- Chen, G. H.; Chen, S.; Liu, Z.; Jiang, F.; and Wang, B. 2024. Humans or LLMs as the Judge? A Study on Judgement Bias. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 8301–8327. Miami, Florida, USA: Association for Computational Linguistics.
- Czarnowska, P.; Vyas, Y.; and Shah, K. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9: 1249–1267.

- Dammu, P. P. S.; Jung, H.; Singh, A.; Choudhury, M.; and Mitra, T. 2024. “They are uncultured”: Unveiling Covert Harms and Social Threats in LLM Generated Conversations. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 20339–20369. Miami, Florida, USA: Association for Computational Linguistics.
- Dennler, N.; Ovalle, A.; Singh, A.; Soldaini, L.; Subramanian, A.; Tu, H.; Agnew, W.; Ghosh, A.; Yee, K.; Peradejordi, I. F.; et al. 2023. Bound by the Bounty: Collaboratively Shaping Evaluation Processes for Queer AI Harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 375–386.
- Dev, S.; Goyal, J.; Tewari, D.; Dave, S.; and Prabhakaran, V. 2023. Building socio-culturally inclusive stereotype resources with community engagement. *Advances in Neural Information Processing Systems*, 36: 4365–4381.
- Dev, S.; Li, T.; Phillips, J. M.; and Srikumar, V. 2020. On measuring and mitigating biased inferences of word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 7659–7666.
- Dev, S.; Sheng, E.; Zhao, J.; Amstutz, A.; Sun, J.; Hou, Y.; Sanseverino, M.; Kim, J.; Nishi, A.; Peng, N.; and Chang, K.-W. 2021. On Measures of Biases and Harms in NLP. *arXiv preprint arXiv:2108.03362*.
- Diehl Martinez, R.; Goriely, Z.; Caines, A.; Buttery, P.; and Beinborn, L. 2024. Mitigating Frequency Bias and Anisotropy in Language Model Pre-Training with Syntactic Smoothing. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 5999–6011. Association for Computational Linguistics.
- D’ignazio, C.; and Klein, L. F. 2020. *Data feminism*. MIT press.
- Ding, L.; Yu, D.; Xie, J.; Guo, W.; Hu, S.; Liu, M.; Kong, L.; Dai, H.; Bao, Y.; and Jiang, B. 2022. Word embeddings via causal inference: Gender bias reducing and semantic information preserving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11864–11872.
- Dwivedi-Yu, J. 2024. FairPair: A Robust Evaluation of Biases in Language Models through Paired Perturbations. In Dinkar, T.; Attanasio, G.; Cercas Curry, A.; Konstas, I.; Hovy, D.; and Rieser, V., eds., *Proceedings of Safety4ConvAI: The Third Workshop on Safety for Conversational AI @ LREC-COLING 2024*, 28–39. Torino, Italia: ELRA and ICCL.
- Friedman, B.; and Nissenbaum, H. 1996. Bias in computer systems. *ACM Transactions on information systems (TOIS)*, 14(3): 330–347.
- Fririksðóttir, S. R.; and Einarsson, H. 2024. Gendered Grammar or Ingrained Bias? Exploring Gender Bias in Icelandic Language Models. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 7596–7610. Torino, Italia: ELRA and ICCL.
- Gaci, Y.; Benatallah, B.; Casati, F.; and Benabdeslem, K. 2022. Debiasing Pretrained Text Encoders by Paying Attention to Paying Attention. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9582–9602. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.
- Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, H.; and Wang, H. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2: 1.
- Gert, B. 2004. *Common Morality: Deciding what to do*. Oxford University Press.
- Ghate, K.; Choudhry, A.; and Bannihatti Kumar, V. 2024. Evaluating Gender Bias in Multilingual Multimodal AI Models: Insights from an Indian Context. In Faleńska, A.; Basta, C.; Costa-jussà, M.; Goldfarb-Tarrant, S.; and Nozza, D., eds., *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 338–350. Bangkok, Thailand: Association for Computational Linguistics.
- Ghosh, S. 2024. Interpretations, Representations, and Stereotypes of Caste within Text-to-Image Generators. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 490–502.
- Ghosh, S.; and Caliskan, A. 2023. ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages. *AAAI/ACM Conference on AI, Ethics, and Society 2023*, 901–912.
- Ghosh, S.; Gautam, S.; Venkit, P.; and Ghosh, A. 2025. Documenting Patterns of Exoticism of Marginalized Populations within Text-to-Image Generators. *arXiv preprint arXiv:2508.02937*.
- Ghosh, S.; Lutz, N.; and Caliskan, A. 2024. “I don’t see myself represented here at all”: User Experiences of Stable Diffusion Outputs Containing Representational Harms across Gender Identities and Nationalities. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 463–475.
- Guan, X.; Demchak, N.; Gupta, S.; Wang, Z.; Ertekin Jr., E.; Koshiyama, A.; Kazim, E.; and Wu, Z. 2025. SAGED: A Holistic Bias-Benchmarking Pipeline for Language Models with Customisable Fairness Calibration. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics*, 3002–3026. Abu Dhabi, UAE: Association for Computational Linguistics.
- Guo, T.; Chen, X.; Wang, Y.; Chang, R.; Pei, S.; Chawla, N. V.; Wiest, O.; and Zhang, X. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.
- Guo, W.; and Caliskan, A. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of*

- the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AAAI/ACM AIES), 122–133.
- Gupta, V.; Venkit, P. N.; Wilson, S.; and Passonneau, R. J. 2023. Sociodemographic Bias in Language Models: A Survey and Forward Path. *arXiv e-prints*, arXiv–2306.
- Hall, S. M.; Gonçalves Abrantes, F.; Zhu, H.; Sodunke, G.; Shtedritski, A.; and Kirk, H. R. 2023. Visogender: A dataset for benchmarking gender bias in image-text pronoun resolution. *Advances in Neural Information Processing Systems*, 36: 63687–63723.
- Han, P.; Kocielnik, R.; Saravanan, A.; Jiang, R.; Sharir, O.; and Anandkumar, A. 2024. ChatGPT Based Data Augmentation for Improved Parameter-Efficient Debiasing of LLMs. In Chakravarthi, B. R.; B, B.; Buitelaar, P.; Durairaj, T.; Kovács, G.; and García Cumberas, M. Á., eds., *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, 73–105. Association for Computational Linguistics.
- Han, X.; Shen, A.; Li, Y.; Fremmann, L.; Baldwin, T.; and Cohn, T. 2022. FairLib: A Unified Framework for Assessing and Improving Fairness. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 60–71. Abu Dhabi, UAE: Association for Computational Linguistics.
- Haraway, D. 1988. ‘Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective’. In *Space, Gender, Knowledge: Feminist Readings*, 53–72. Routledge.
- He, Z.; Wang, Y.; McAuley, J.; and Majumder, B. P. 2022. Controlling Bias Exposure for Fair Interpretable Predictions. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2022*, 5854–5866. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.
- Herold, B.; Waller, J.; and Kushalnagar, R. 2022. Applying the stereotype content model to assess disability bias in popular pre-trained NLP models underlying AI-based assistive technologies. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, 58–65.
- Hosseini, S.; Palangi, H.; and Awadallah, A. H. 2023. An Empirical Study of Metrics to Measure Representational Harms in Pre-Trained Language Models. In Ovalle, A.; Chang, K.-W.; Mehrabi, N.; Pruksachatkun, Y.; Galystan, A.; Dhamala, J.; Verma, A.; Cao, T.; Kumar, A.; and Gupta, R., eds., *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, 121–134. Toronto, Canada: Association for Computational Linguistics.
- Huang, Y.; and Xiong, D. 2024. CBBQ: A Chinese Bias Benchmark Dataset Curated with Human-AI Collaboration for Large Language Models. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2917–2929. Torino, Italia: ELRA and ICCL.
- Jeyaraj, M.; and Delany, S. 2024. An Explainable Approach to Understanding Gender Stereotype Text. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 45–59.
- Jha, A.; Prabhakaran, V.; Denton, R.; Laszlo, S.; Dave, S.; Qadri, R.; Reddy, C.; and Dev, S. 2024. ViSAGE: A Global-Scale Analysis of Visual Stereotypes in Text-to-Image Generation. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12333–12347. Bangkok, Thailand: Association for Computational Linguistics.
- Ji, C.; Zhao, B.; Wang, Z.; Wang, Y.; Zhang, Y.; Cheng, Y.; Feng, R.; and Zhang, X. 2025. RoBGuard: Enhancing LLMs to Assess Risk of Bias in Clinical Trial Documents. In *Proceedings of the 31st International Conference on Computational Linguistics*, 1258–1277.
- Kaneko, M.; Imankulova, A.; Bollegala, D.; and Okazaki, N. 2022. Gender bias in masked language models for multiple languages. *Proceedings of the 2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.
- Kay, J.; Kasirzadeh, A.; and Mohamed, S. 2024. Epistemic injustice in generative AI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 684–697.
- Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. S. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, 2564–2572. PMLR.
- Kirk, H. R.; Jun, Y.; Volpin, F.; Iqbal, H.; Benussi, E.; Dreyer, F.; Shtedritski, A.; and Asano, Y. 2021. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Advances in neural information processing systems*, 34: 2611–2624.
- Kitchenham, B. 2004. Procedures for Performing Systematic Reviews. 33.
- Knopf, J. W. 2006. Doing a literature review. *PS: Political Science & Politics*, 39(1): 127–132.
- Kohn, S. 2013. Affirmative Action Has Helped White Women More Than Anyone. *TIME Magazine*.
- Leidinger, A.; and Rogers, R. 2024. How are LLMs mitigating stereotyping harms? Learning from search engine studies. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 839–854.
- Li, L.; Chen, J.; Lu, X.; Lu, Y.; Lin, H.; Zhou, S.; Zhu, H.; Wang, W.; Liu, Z.; Han, X.; et al. 2024. Debiasing In-Context Learning by Instructing LLMs How to Follow Demonstrations. In *Findings of the Association for Computational Linguistics ACL 2024*, 7203–7215.
- Li, Y.; Zhang, G.; Yang, B.; Lin, C.; Ragni, A.; Wang, S.; and Fu, J. 2022. HERB: Measuring Hierarchical Regional Bias in Pre-trained Language Models. In He, Y.; Ji, H.; Li, S.; Liu, Y.; and Chang, C.-H., eds., *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, 334–346. Online only: Association for Computational Linguistics.

- Luccioni, S.; Akiki, C.; Mitchell, M.; and Jernite, Y. 2024. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, 36.
- Luo, H.; and Glass, J. 2023. Logic Against Bias: Textual Entailment Mitigates Stereotypical Sentence Reasoning. In Vlachos, A.; and Augenstein, I., eds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 1243–1254. Dubrovnik, Croatia: Association for Computational Linguistics.
- Mack, A.; Qadri, R.; Denton, R.; Kane, S. K.; and Bennett, C. L. 2024. “They only care to show us the wheelchair”: Disability Representation in text-to-image AI models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.
- Mack, K. A.; McDonnell, E.; Jain, D.; Lu Wang, L.; E. Froehlich, J.; and Findlater, L. 2021. What do we mean by “accessibility research”? A literature survey of accessibility papers in CHI and ASSETS from 1994 to 2019. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–18.
- Marchiori Manerba, M.; Stanczak, K.; Guidotti, R.; and Augenstein, I. 2024. Social Bias Probing: Fairness Benchmarking for Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 14653–14671. Miami, Florida, USA: Association for Computational Linguistics.
- Meade, N.; Poole-Dayana, E.; and Reddy, S. 2022. An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1878–1898. Dublin, Ireland: Association for Computational Linguistics.
- Mei, K.; Fereidooni, S.; and Caliskan, A. 2023. Bias against 93 stigmatized groups in masked language models and downstream sentiment classification tasks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1699–1710.
- Miceli, M.; Posada, J.; and Yang, T. 2022. Studying up machine learning data: Why talk about bias when we mean power? *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP): 1–14.
- Mitchell, M.; Wu, S.; Zaldívar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, 220–229.
- Mujtaba, D.; Mahapatra, N.; Arney, M.; Yaruss, J.; Gerlach-Houck, H.; Herring, C.; and Bin, J. 2024. Lost in Transcription: Identifying and Quantifying the Accuracy Biases of Automatic Speech Recognition Systems Against Disfluent Speech. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 4795–4809. Mexico City, Mexico: Association for Computational Linguistics.
- Olteanu, A.; Castillo, C.; Diaz, F.; and Kiciman, E. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data*, 2: 13.
- Online Etymology Dictionary. 2025. bias, n. In *Online Etymology Dictionary*, 5412.
- Oxford English Dictionary. 2025. bias, n. In *Oxford English Dictionary Online*, 1530.
- Oxford English Dictionary Online. 2025. bias, v. In *Oxford English Dictionary Online*, 1532.
- Pang, R. Y.; Schroeder, H.; Smith, K. S.; Barocas, S.; Xiao, Z.; Tseng, E.; and Bragg, D. 2025. Understanding the LLMification of CHI: Unpacking the Impact of LLMs at CHI through a Systematic Literature Review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713941.
- President of the United States. 2025. Executive Order 14179: Advancing U.S. Leadership in Artificial Intelligence.
- Qadri, R.; Shelby, R.; Bennett, C. L.; and Denton, E. 2023. AI’s Regimes of Representation: A Community-centered Study of Text-to-Image Models in South Asia. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 506–517.
- Raghavan, M.; Barocas, S.; Kleinberg, J.; and Levy, K. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 469–481.
- Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 33–44.
- Ranaldi, L.; Ruzzetti, E. S.; Venditti, D.; Onorati, D.; and Zanzotto, F. M. 2024. A Trip Towards Fairness: Bias and De-Biasing in Large Language Models. In Bollegala, D.; and Shwartz, V., eds., *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, 372–384. Mexico City, Mexico: Association for Computational Linguistics.
- Sap, M.; Gabriel, S.; Qin, L.; Jurafsky, D.; Smith, N. A.; and Choi, Y. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5477–5490. Online: Association for Computational Linguistics.
- Saunders, D.; and Byrne, B. 2020. Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7724–7736. Online: Association for Computational Linguistics.

- Sharma, S.; Kulkarni, A.; Suresh, T.; Mathur, H.; Nakov, P.; Akhtar, M. S.; and Chakraborty, T. 2023. Characterizing the Entities in Harmful Memes: Who is the Hero, the Villain, the Victim? In Vlachos, A.; and Augenstein, I., eds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2149–2163. Association for Computational Linguistics.
- Shin, J.; Song, H.; Lee, H.; Jeong, S.; and Park, J. 2024. Ask LLMs Directly, “What shapes your bias?”: Measuring Social Bias in Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 16122–16143. Bangkok, Thailand: Association for Computational Linguistics.
- Snyder, H. 2019. Literature review as a research methodology: An overview and guidelines. *Journal of business research*, 104: 333–339.
- Srinivasan, T.; and Bisk, Y. 2022. Worst of Both Worlds: Biases Compound in Pre-trained Vision-and-Language Models. In Hardmeier, C.; Basta, C.; Costa-jussà, M. R.; Stanovsky, G.; and Gonen, H., eds., *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 77–85. Seattle, Washington: Association for Computational Linguistics.
- Stacey, J.; Minervini, P.; Dubossarsky, H.; Riedel, S.; and Rocktäschel, T. 2020. Avoiding the Hypothesis-Only Bias in Natural Language Inference via Ensemble Adversarial Training. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8281–8291. Association for Computational Linguistics.
- Steed, R.; and Caliskan, A. 2021. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 701–713.
- Suresh, H.; and Gutttag, J. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and access in algorithms, mechanisms, and optimization*, 1–9.
- Tan, Y. C.; and Celis, L. E. 2019. Assessing social and intersectional biases in contextualized word representations. *Advances in neural information processing systems*, 32.
- Toro Isaza, P.; Xu, G.; Oloko, T.; Hou, Y.; Peng, N.; and Wang, D. 2023. Are Fairy Tales Fair? Analyzing Gender Bias in Temporal Narrative Event Chains of Children’s Fairy Tales. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6509–6531. Association for Computational Linguistics.
- Trisovic, A.; Lau, M. K.; Pasquier, T.; and Crosas, M. 2022. A large-scale study on research code quality and execution. *Scientific Data*, 9(1): 60.
- Ueda, N.; Mita, M.; Oka, T.; and Komachi, M. 2024. Token-length Bias in Minimal-pair Paradigm Datasets. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 16224–16236.
- Varner, M.; and Lecher, C. 2023. How We Investigated L.A.’s Homelessness Scoring System. *The Markup*.
- Venkit, P.; Srinath, M.; Gautam, S.; Venkatraman, S.; Gupta, V.; Passonneau, R. J.; and Wilson, S. 2023. The Sentiment Problem: A Critical Survey towards Deconstructing Sentiment Analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 13743–13763.
- Venkit, P. N.; Srinath, M.; and Wilson, S. 2022. A study of implicit bias in pretrained language models against people with disabilities. In *Proceedings of the 29th International Conference on Computational Linguistics*, 1324–1332.
- Venkit, P. N.; Srinath, M.; and Wilson, S. 2023. Automated Ableism: An Exploration of Explicit Disability Biases in Sentiment and Toxicity Analysis Models. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, 26–34.
- Wan, Y.; Pu, G.; Sun, J.; Garimella, A.; Chang, K.-W.; and Peng, N. 2023. ‘Kelly is a Warm Person, Joseph is a Role Model’: Gender Biases in LLM-Generated Reference Letters. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 3730–3748. Singapore: Association for Computational Linguistics.
- Wan, Y.; Subramonian, A.; Ovalle, A.; Lin, Z.; Suvarna, A.; Chance, C.; Bansal, H.; Pattichis, R.; and Chang, K.-W. 2024. Survey of bias in text-to-image generation: Definition, evaluation, and mitigation. *arXiv preprint arXiv:2404.01030*.
- Wolfe, R.; Dangol, A.; Howe, B.; and Hiniker, A. 2024. Representation Bias of Adolescents in AI: A Bilingual, Bicultural Study. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 1621–1634.
- Wu, Y.; Nakashima, Y.; and Garcia, N. 2024. Stable diffusion exposed: Gender bias from prompt to image. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 1648–1659.
- Yifei, L.; Ungar, L.; and Sedoc, J. 2023. Conceptor-Aided Debiasing of Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 10703–10727. Singapore: Association for Computational Linguistics.
- You, Z.; Lee, H.; Mishra, S.; Jeoung, S.; Mishra, A.; Kim, J.; and Diesner, J. 2024. Beyond Binary Gender Labels: Revealing Gender Bias in LLMs through Gender-Neutral Name Predictions. In Faleńska, A.; Basta, C.; Costa-jussà, M.; Goldfarb-Tarrant, S.; and Nozza, D., eds., *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 255–268. Bangkok, Thailand: Association for Computational Linguistics.
- Yun, H. S.; Marshall, I. J.; Trikalinos, T. A.; and Wallace, B. C. 2023. Appraising the potential uses and harms of LLMs for medical systematic reviews. In *Proceedings of*

the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), 10122—10139.

Zeng, H.; Yue, Z.; Zhang, Y.; Shang, L.; and Wang, D. 2024. Fair federated learning with biased vision-language models. In *Findings of the Association for Computational Linguistics ACL 2024*, 10002–10017.

Zhang, H.; Sneyd, A.; and Stevenson, M. 2020. Robustness and Reliability of Gender Bias Assessment in Word Embeddings: The Role of Base Pairs. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 759–769. Association for Computational Linguistics.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In Palmer, M.; Hwa, R.; and Riedel, S., eds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2979–2989. Copenhagen, Denmark: Association for Computational Linguistics.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).