Metrics for What, Metrics for Whom: Assessing Actionability of Bias Evaluation Metrics in NLP

Pieter Delobelle[1*], Giuseppe Attanasio[2*], Debora Nozza[3], Su Lin Blodgett[4], Zeerak Talat[5]

[1]KU Leuven; Leuven.ai, [2]Instituto de Telecomunicações, Lisbon, [3]MilaNLP, Bocconi
[4]Microsoft Research Montréal, [5]Mohamed bin Zayed University of Artificial Intelligence

Abstract

This paper introduces the concept of *actionability* in the context of bias measures in natural language processing (NLP). We define actionability as the degree to which a measurement's results enable informed action and propose a set of desiderata for assessing it. Building on existing frameworks such as measurement modeling, we argue that actionability is a crucial aspect of bias measures that has been largely overlooked in the literature. We conduct a comprehensive review of 146 papers proposing bias measures in NLP, examining whether and how they provide the information required for actionable results. Our findings reveal that many key elements of actionability, including a measure's intended use and reliability assessment, are often unclear or absent. This study highlights a significant gap in the current approach to developing and reporting bias measures in NLP. We argue that this lack of clarity may impede the effective implementation and utilization of these measures. To address this issue, we offer recommendations for more comprehensive and actionable metric development and reporting practices in NLP bias research.

1 Introduction

As the landscape of bias measures in natural language processing (NLP) has expanded, so too has the literature examining and interrogating these measures (e.g., Blodgett et al., 2021; Goldfarb-Tarrant et al., 2021; Delobelle et al., 2022; Orgad and Belinkov, 2022; Selvam et al., 2023; Goldfarb-Tarrant et al., 2023c; Tokpo et al., 2023). In particular, increasingly rich reflections within and beyond NLP have offered vocabularies and frameworks for navigating this landscape; for example, the framework of *measurement modeling* from the quantitative social sciences disentangles what is measured (a theoretical construct) from how it is measured (its operationalization), and offers the vocabulary of *validity* and *reliability* for assessing measures (Jacobs and Wallach, 2021; Blodgett et al., 2021).

Across the literature proposing and examining bias measures, talk about measures is often informally tied to talk about what can be done with results produced by measures—i.e., measures' results are often used in decision-making, and good measures should not only exhibit characteristics such as validity and reliability, but should also facilitate decision-making or intervention. For example, natural language generation practitioners use the results of automated metrics to select which models should undergo human evaluation (Zhou et al., 2022b), while other measures' results might guide policies for model release and deployment (Solaiman, 2023). Together, this suggests another piece of vocabulary with which we might assess bias measures. In this paper, we seek to formalize this intuition by introducing *actionability*—the degree to which a measure's results enable informed action—and outlining a set of *desiderata* for actionability—what information is required of a bias measure in order to act based on its results.

At the same time, while the measurement modeling literature has shown the importance of clearly conceptualizing bias and establishing bias measures' validity and reliability, it has also shown that the NLP literature routinely fails to do so. For example, bias in the NLP literature is often underspecified (Blodgett et al., 2020), and measures are often poorly matched to the constructs they are intended to measure (Gonen and Goldberg, 2019; Blodgett et al., 2021) or lack sufficient description to establish a match altogether (Goldfarb-Tarrant et al., 2023c). Hypothesizing that the literature may similarly seldom assess what bias measures can be used for, and whether enough information is provided to facilitate that use, we conduct a review of 146 papers proposing bias measures, examining whether and how papers provide the information required to act based on the proposed measures' results.

[*]Joint first authors.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 21669–21691

We find that many desiderata for actionability, such as a bias measure's intended use or an assessment of its reliability, are often not clearly provided or go unstated altogether. We argue that this lack of clear information may hinder bias measures' effective implementation and use, and offer suggestions for improving the development and dissemination of bias measures in NLP research.

2 Actionability

In this section, we introduce and formalize actionability, draw connections between actionability and other concepts related to the trustworthy NLP literature, and provide an example of a bias measure and the actions it facilitated.

We introduce actionability in order to answer the question: What is required of a bias measure in order to take informed actions based on its results? Following Dev et al. (2022), we define a bias measure as an "evaluation standard that includes a metric(s) applied to a dataset" which is applied to measure "bias," itself a contested and often underspecified construct (Blodgett et al., 2020). Throughout the paper we use "bias" to refer expansively to the wide range of concerns, impacts, and harms that work in the NLP literature has sought to measure under the term "bias."

Actionability refers to the degree to which a measure's results enable decision-making or intervention; that is, results from actionable bias measures should facilitate informed actions with respect to the bias under measurement. Such results might communicate aspects of the measured bias such as who is impacted or harmed by a system, the degree and scale of impact or harm, or potential sources of the issue. In turn, the decisions or interventions that these results enable might include targeted improvements to training or fine-tuning processes (e.g., Talat and Lauscher, 2022; Lauscher et al., 2021; Delobelle and Berendt, 2023; Bartl et al., 2020; Attanasio et al., 2022), deployment of appropriate safeguards (e.g., Tamkin et al., 2023; Suau et al., 2024; Bauer et al., 2024), decisions to re-design or not to deploy (Birhane et al., 2024), or changes in regulation or policy (Kolkman, 2020; Sztandar-Sztanderska and Zieleńska, 2022).

The ability to act on bias measure results may not be equally distributed among stakeholders, as power or organizational dynamics can shape their ability to intervene. For example, while some results may suggest that retraining a model or delaying a system's deployment would be effective interventions, stakeholders might not be equally empowered to take such actions. Stakeholders such as consumers may only be in a position to opt out of using or providing data for a system (Gangadharan, 2021), while regulators may choose to sanction—e.g., by issuing fines or outright banning uses that are not compliant with regulation—or allow particular applications.

To better situate actionability, we consider it against other concepts the responsible NLP literature—specifically, accountability, transparency, interpretability, and validity—beginning with **accountability**. Evaluations or audits of AI systems are often conducted (implicitly or explicitly) with the goals of "establish[ing] informed and consequential judgments of... AI systems" (Birhane et al., 2024)—e.g., whether a system's behavior is legally compliant—and holding AI providers accountable—i.e., "responsible or answerable for a system, its behavior and its potential impacts" (Raji et al., 2020). However, as Birhane et al. write, in practice "AI audit studies do not consistently translate into more concrete objectives to regulate system outcomes." Thus, we see the actionability of a bias measure as a component for ensuring that results from bias measures can translate into action that shapes system outcomes and policy and holds providers responsible.

Research on the transparency of AI systems has argued for the importance of "develop[ing] more trustworthy AI" (Larsson and Heintz, 2020). Using Liao and Wortman Vaughan's (2024) definition of (informational) **transparency** as "what information about a model [or system] should be disclosed to enable appropriate understanding," we see transparency as required for actionability. That is, it is impossible for stakeholders to act upon the results of a bias measure without crucial knowledge about a system's design and deployment.

The **interpretability** of a model attends to whether we understand the process by which the model produces an output (Ribeiro et al., 2016; Doshi-Velez and Kim, 2017). Analogously, the interpretability of a bias measure attends to whether we understand the process by which a bias measure arrives at a result. Unlike actionability, interpretability does not attend to whether that result enables informed interventions—even if understanding why a measure produces a result can help facilitate interventions (Attanasio et al., 2023).

The aspects of **validity** most closely related to

actionability are the three addressing the use and utility of a measure's results: consequential validity, predictive validity, and hypothesis validity.[1] **Consequential validity** involves "identifying and evaluating the consequences of using the measurements obtained from a measurement model" (Jacobs and Wallach, 2021). For bias measures, for example, using a certain measure may make harm to some populations more visible than others, depending on which populations the measure was designed for, or else a measure's uptake—and the subsequent optimization of models towards it—may have unintended effects. Thus, consequential validity is related to actionability, as one consequence of using a bias measure's results is precisely the decisions or interventions that might be made on the basis of those results; therefore, in developing actionable metrics, practitioners should consider the consequences of the decisions and interventions that those metrics facilitate.

Meanwhile, **predictive validity** captures "the extent to which measurements obtained from a measurement model are predictive of measurements of any relevant observable properties... thought to be related to the construct purported to be measured," while **hypothesis validity** captures "the extent to which the measurements obtained from a measurement model support substantively interesting hypotheses about the construct purported to be measured" (Jacobs and Wallach, 2021).[2] We argue that for bias measures, actionability is very closely related to predictive and hypothesis validity, as bias measure results that enable decisions or interventions also implicitly or explicitly support a particular type of hypothesis—i.e., a hypothesis that some decision(s) or intervention(s) can meaningfully address the bias under measurement. While actionability can thus be understood as a narrower form of hypothesis validity, we propose it as its own concept to draw attention to the specific types of hypotheses—i.e., about meaningful decisions or interventions—that we argue bias measures should support.

While other types of validity—face, content, convergent, and divergent validity—appear less directly related to actionability conceptually, we see them as no less important; bias measures that do not capture all relevant aspects of the bias to be measured, or whose results are implausible or fail to correlate with other measures' results (Jacobs and Wallach, 2021), are unlikely to enable informed action.

Similarly, measures that are not **reliable** are unlikely to be actionable, as their results may not provide a sufficient basis for making well-informed decisions. In this desideratum we include test-retest reliability—i.e., whether similar inputs yield similar results (Jacobs and Wallach, 2021)—as well as the reporting of a measure's margins of error; statistical tests used to assess results' significance (Goodman et al., 2016); and other analyses of possible sources of uncertainty of results (Barrainkua et al., 2023; Black et al., 2024), such as variation due to choices of seed words (Antoniak and Mimno, 2021) or templates (Delobelle et al., 2022).

Example. In 2014, Amazon sought to develop an AI system for screening candidate resumés, which was ultimately discontinued in 2018 because it ranked female candidates lower than male candidates (Anonymous, 2016). While we do not know the exact details of the bias measure(s) Amazon used to assess the system, the results did facilitate understanding of who might have been impacted—people who had attended women's colleges used the word "women's" on their resumés, or did not use words "more commonly found on male engineers" resumés, such as 'executed"'—all disproportionately women and gender minorities (Dastin, 2018). We also know that the results enabled at least three actions: first, Amazon attempted to mitigate the issue, "edit[ing] the programs to make them neutral to [the terms mentioned]"; second, Amazon discontinued the use of the system for ranking candidates and "disbanded the team [building the system]"; and finally, Amazon moved towards using a "'much-watered down version"' for to help with 'rudimentary chores,' including culling duplicate candidate profiles" (Dastin, 2018).

This example illustrates how results from bias measures can facilitate various actions from various stakeholders, including mitigation attempts by system developers and decisions to discontinue or to use alternate versions for different purposes by (presumably) Amazon leadership. It further illustrates the importance of transparency—specifically, the lack of external transparency with respect to results of Amazon's bias measures, and to the active use of the system between 2015 and 2018. Had

[1] Validity has been conceptualized in several ways; we use the conceptualization from Jacobs and Wallach (2021).

[2] We consider predictive and hypothesis validity together because, as Jacobs and Wallach (2021) point out, "the main distinction between predictive validity and hypothesis validity hinges on the definition of 'substantively interesting hypotheses,"' and that "distinction is not always clear cut."

the biases of the system been public knowledge, stakeholders outside the project team and Amazon leadership would have been able take action—e.g., individuals would have been able to withdraw applications or choose not to apply, while regulators would have the ability to sanction the use of a system that was in breach of regulation around gender discrimination in hiring. Insofar that the results of a bias measure of the system are not disclosed to the public and regulators, both are precluded from informed and meaningful action.

3 Desiderata for Actionability

What, concretely, makes a bias measure actionable? In this section, we outline desiderata for bias measures—i.e., information that a measure should provide and justify to enhance its actionability. We draw these desiderata from prior literature related to responsible NLP, including work on fairness in machine learning and NLP (Mitchell et al., 2021; Czarnowska et al., 2021), measurement (Blodgett et al., 2021; Jacobs and Wallach, 2021), and AI auditing and algorithmic accountability (Raji et al., 2020; Birhane et al., 2024). We will also use these desiderata as the basis for our taxonomy and survey in the remainder of this paper.

Motivation. The motivation for a proposed bias measure specifies what *need* the measure is intended to address, e.g., measuring direct discrimination (Sweeney and Najafian, 2019), adapting to new socio-cultural contexts (Bhatt et al., 2022), or extending to new languages (Huang et al., 2020).

A clearly described motivation can increase a measure's actionability by helping people using the measure to assess whether the bias they seek to measure and the system and context of use for which they seek to measure bias are well-matched to the need the measure is intended to address.

Underlying bias construct. Drawing on measurement modeling, we view bias as an unobservable theoretical construct operationalized via bias measures (Jacobs and Wallach, 2021). Under this view, a proposed bias measure is always accompanied, implicitly or explicitly, by an underlying theoretical understanding of what constitutes bias. However, these theoretical understandings are not always clearly specified or conceptualized; Blodgett et al. (2020) illustrate that "bias" in the NLP literature is often underspecified, and Jacobs and Wallach (2021) argue that disagreements in the AI fairness literature often arise because authors rarely make explicit their theoretical understandings of fairness, which has many "context-dependent, and sometimes even conflicting" understandings.

We argue that clarity in the conceptualization of a bias measure's underlying construct can increase the measure's actionability, as a bias construct articulates the measure's scope—e.g., what impacts or harms the measure is intended to capture, for which populations those impacts or harms are intended to be captured, or what constitutes impact or harm. If the bias construct is not clearly specified and conceptualized, it becomes unclear how the measure's results speak to any impacts or harms, and is therefore unlikely that those results can facilitate informed action.

Interval and ideal result. Understanding, and therefore acting, on the results of a measure requires clearly articulated information about the values a result can take on; these values inform the statistical analyses that can be performed and the interpretations that can be made. Minimally, actionability requires descriptions of: first, the numerical domain of the result (natural, real, or rational);[3] second, the interval a measure operates on—i.e., the values the result can take on—which may or may not be bounded (log-likelihood-based measures being an example of the latter (Webster et al., 2021)); and third, the scale of the interval—for example, for measures on a logarithmic scale a result of 10 might be much worse than 3, but not that much better than 20. The numerical domain and the bounds and scale of the interval are necessary for interpreting the result, as they allow people using the measure to estimate how far the result is from the interval bounds and what it might mean relative to other possible results.

Proposed bias measures should also specify an ideal result. The choice of an ideal result is inherently normative, as it reflects a measure creator's perspective on what constitutes desired system behavior and how that is expressed in the measure's result.[4] Specifying an ideal result can facilitate a measure's actionability by providing people using the measure with a clear goal or requirement, particularly if the choice is explicitly connected to the underlying bias construct and its wider socio-

[3]The interval might also not exist, e.g., in the case of results taking on binary values.

[4]Setting ideal scores can be a difficult task that requires taking into account social context, risks and desired outcomes. See Kearns et al. (2018) for further discussion.

historical context—e.g., for hiring, an ideal result might be adherence to the four-fifths rule, a guideline for assessing what constitutes discrimination in employment in the U.S. (Ajunwa et al., 2016).[5]

Intended use. Proposed bias measures should specify under what circumstances or conditions the measure should be expected to produce meaningful results. This can include, for example, what types of models or additional data are required to be used in conjunction with the measure or which hyper-parameters govern the behavior of the measure. Broadly, intended use seeks to describe a wide variety of conditions that may be mechanistic—-e.g., models, data, or hyper-parameters—or social, e.g., particular social settings in which the result of a measure is considered meaningful.

For example, in some measures, the metric is closely tied to a specific dataset or dataset format—e.g., StereoSet's (Nadeem et al., 2021) stereotyping score aggregates a model's preferences for stereotypical versus anti-stereotypical completions and therefore requires a dataset containing such stereotype/anti-stereotype pairs. Moreover, StereoSet's particular construction—i.e., its use of log-likelihood and pseudo-perplexity to measure stereotyping—are designed for use with masked language models and auto-regressive language models respectively. By contrast,CrowS-Pairs (Nangia et al., 2020), a similar measure, only uses pseudo-perplexity and can therefore only be used with masked language models.Thus, the construction of StereoSet and CrowS-Pairs limits their applicability to certain dataset and model characteristics, and they may therefore be poorly matched with other settings.

Providing descriptions of the mechanical conditions and socio-historical context and that render the result of a measure meaningful facilitates actionability by bounding a measure's application space, thereby giving potential users of a measure the information needed to assess whether the measure is appropriate for their use cases. In particular, when metrics and datasets are introduced together to propose a new measure, specifying the intended use can help to clarify how the dataset and metric together make the measure fit-for-purpose, as well as what other data the metric might potentially be appropriately applied to, and vice versa.

[5]Ideal results are also often used in the standards identification phase of AI audits (Birhane et al., 2024), to "effectively articulat[e] the requirements for an ideal AI audit outcome."

Reliability. As we discusss in Section 2, we view the reliability of a bias measure as a prerequisite for actionability. Thus, proposed bias measures should explain how their reliability was assessed.

4 Literature review and analysis

To identify current trends and existing gaps in the field, we conduct a literature review, examining how papers proposing bias metrics engage with our desiderata for actionability. While previous studies (e.g., Blodgett et al., 2020; Sheng et al., 2021; Goldfarb-Tarrant et al., 2023c; Liu et al., 2023) have explored how responsible NLP concerns (including bias) and measures of those concerns are described in the NLP literature, to the best of our knowledge, this is the first review specifically focused on the actionability of bias measures.

Search methodology. Our search and paper selection processes follow the PRISMA 2020 guidelines (Page et al., 2021) for systematic reviews and meta-analyses (see Figure 1 in Appendix A for an overview diagram).

We used the ACL Anthology API to identify all papers whose title or abstract contains at least one of the keywords "fair," "bias," or "stereotyp*" and which co-occur with either "eval*" or "metric."[6] Our search included all work published before April 2024. We augmented the initial set by adding four papers from Delobelle et al. (2022) and one paper from Orgad and Belinkov (2022), two comprehensive surveys of recent bias evaluation approaches. This yielded a total of 1181 papers.

Paper selection. Two of the authors filtered the papers for relevance by reading titles and abstracts,[7] removing papers not written in English or not proposing a new bias measure. As we describe in Section 2, we define a bias measure as an "evaluation standard that includes a metric(s) applied to a dataset" (Dev et al., 2022) which is applied to measure bias. We use an intentionally expansive definition to include a wide range of measures for a wide range of biases to capture as broad a view of the literature as possible.

The two authors conducting the screening initially examined a shared pool of 140 papers, yield-

[6]We acknowledge that there might be papers that introduce bias metrics for NLP models *outside* of the ACL community. See Limitations (§7) for a discussion.

[7]If the title and abstract did not provide sufficient details to decide, we read the full paper. In the few cases where it was not readily apparent whether a paper introduced a new metric, the authors all met to discuss the paper.

ing an inter-annotator Fleiss kappa of $\kappa = 0.76$. Disagreements during this initial screening arose due to lack of clarity with respect to several inclusion criteria, including what constitutes "bias" (e.g., caricatures (Cheng et al., 2023)) and a new measure (i.e., a new dataset, a new metric, or both). After discussion among the authors, we chose to resolve these as expansively as possible: we include any papers that self-describe as engaging with bias or stereotyping, regardless of how those terms are conceptualized, and we included not only papers introducing both a new dataset and a new metric but also papers introducing just one or the other—e.g., a paper adapting a measure from one language variety to another by introducing a dataset in the second language variety, to which the original metric is intended to be applied. The authors then screened the full set of 1181 papers, obtaining a final set of 146 papers.[8]

Annotation. We annotate each paper in our final set for whether and how it provides the information required by our desiderata for actionability (Section 3). Nearly all of the desiderata require open-ended descriptions—e.g., of the bias construct to be measured. We annotate for each desideratum by extracting all directly relevant passages—e.g., the passage(s) describing the bias construct—noting if no passages match. For the ideal result and reliability desiderata, we extracted two binary values: whether each is described in the paper, and if so whether each was clearly justified or assessed.

5 Threats to Actionability

Measures' stated motivations rarely linked to their use. We read and categorized all free-form text passages describing motivations into a categorical schema using an inductive process. For 20% of papers, we were unable to identify any text passage with a clear motivation for introducing a new measure. In all other cases, we were able to identify clear motivations such as extending the measure to another language, setting, or modality. A subset of papers providing a motivation are motivated by improving existing measures, e.g., Dinan et al.'s (2020b) measure that "allow(s) for better identification of gender bias," or by addressing reliability or reproducibility concerns.

[8]From a qualitative analysis, among excluded papers we found *i*) papers mentioning inductive, lexical, or syntactic bias, and *ii*) other papers related to social bias that that did not introduce a bias measure, e.g., debiasing methods.

Although 80% of the papers provide a motivation, the degree to which that motivation is clear and specific varies, leaving a large subset of papers either vaguely gesturing towards a motivation. For example, Yeh et al. (2023) motivate their work on measuring bias in LLMs due to the existence of "LangChain," an underspecified "threat."

> "Although a plethora of research has been dedicated to identifying bias in LLMs and formulating debiasing techniques, there remains an under-examined threat capable of directly impacting LLMs using external data without necessitating significant computational training resources. This hazard is termed 'LangChain."' – Yeh et al. (2023)

Similarly, while introducing a new debiasing method for contextualized representations, Basu Roy Chowdhury et al. (2021) introduce the use of MDL as a bias measure as it is "finer grained," however it is unclear why the granularity of accuracy is unsatisfactory in their use case, or why other measures, e.g., non-probing-based methods, were not considered.

> "We extend previous evaluation methodology for debiasing by measuring Minimum Description Length (MDL) [...] of labels given representations, instead of probing accuracy. MDL provides a finer-grained evaluation benchmark for measuring debiasing performance." – Basu Roy Chowdhury et al. (2021)

Vague or non-existing motivations present a barrier to the use of a measure, as readers are forced to infer what need or use case a measure addresses and whether a measure is appropriate for their use case. Providing a clear motivation can be a simple task. For instance, papers might introduce aspects of bias that are not represented in other measures but which they argue to be important:

> "However, *one aspect of bias that has received less attention is offensive stereotyping toward marginalised groups*. For example, using slurs to describe non-white or LGBTQ communities or using swear words to describe women." – Elsafoury et al. (2022)

Even for papers that do provide a concrete motivation (see Table 1 for a breakdown), those motivations are routinely disconnected from the measures

that are ultimately proposed. For example, Li et al. (2022) motivate the work by referring to allocational harms in a resume classification system:

> "Bias in NLP applications makes distinct judgements 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)

However, their measure quantifies stereotypical group representations instead of the performance differences or impacts on job seekers that this motivation alludes to.

Missing construct definition. For 25% of papers, it was impossible to understand what theoretical bias construct the authors intended to measure. For these papers, we were either unable to identify a text passage describing the underlying construct, or the construct definition was highly underspecified—e.g., "immigrant bias" (Goldfarb-Tarrant et al., 2023b). This finding is particularly surprising considering recent critiques. For example, about one-third of this set of papers cite Blodgett et al. (2020) explicitly, who argued for the importance of clearly defining "bias." As 72% of all papers in our sample were published after 2020. We therefore echo the argument presented by Blodgett et al. in 2020: that without a well-defined theoretical bias construct "techniques are poorly matched to their motivations, and are not comparable to one another", and that without a well-defined theoretical bias construct, assessing the match between construct definition and operationalization is impossible. Moreover, we believe that such a lack forecloses meaningful analysis or action on the basis of a measure's result.

On a more positive note, we observe that 36% of the papers include an explicit "Bias statement" (Hardmeier et al., 2021). Such statements range from brief descriptions relying on existing literature (e.g., Jeoung et al. (2023), or on theory about stereotyping developed in Fiske (2018)), to more detailed descriptions (e.g., Malik et al.'s (2022) explanation of the caste system in India). Another 15% of the papers discuss downstream harms and the risks of biased behaviors that the proposed metric is intended to capture.

Mismatch between construct and its operationalization. We found that in 24% of the papers the theoretical bias construct and operationalization choices for the metric are conflated. Most often, these papers do not discuss an underlying construct and instead rely on other bias measures—often WEAT (Caliskan et al., 2017)—to define "bias." Such choices, omitting a description or conflating the definition and operationalization, present challenges to actionability. Similarly, we identified instances where the construct and the operationalization were not aligned. For example, España-Bonet and Barrón-Cedeño (2022) (vaguely) conceptualized bias as *social cultural biases*, including racism, ageism, sexism. Then, they operationalize the measure using a WEAT test (Caliskan et al., 2017). However, they measure two WEAT tests[9] that are unrelated to the described bias construct

Reporting of interval and ideal result. Our analysis shows most papers (82%) report an interval or variation of the measure they propose. Of these papers, 58% use a bounded range (e.g., $[-1, 1]$, $[0, 1]$), or their percentage equivalents. Other papers (12) use logarithmic or other operators thatwresult in unbounded intervals on one or both sides.[10] However, even when evaluated against a reference, the relationship between the score and the impacts, e.g., the amount of stereotypical associations made in generated text, remains opaque. It is, therefore, necessary to measure against some external reference which is grounded in measuring the severity of a model's generations or predictions.

Many papers (77%) explicitly indicate the ideal result a model should attain when assessed with their proposed measure. Yet, only 32% of those papers engage in discussions around the ideal result or offer insights into its interpretation. One method for discussing the ideal outcome is to explicitly describe the behaviour of an 'ideal' model, e.g.,

> "**IdealLM** We define this hypothetical model as the one that always *picks correct associations for a given target term context*. It also *picks equal number of stereotypical and anti-stereotypical associations over all the target terms*. So the resulting lms and ss scores are 100 and 50 respectively." – Nadeem et al. (2021)

[9] Pleasant/unpleasant versus flowers/insects (WEAT1) and musical instruments/weapons (WEAT2).

[10] Unbounded intervals are often a natural consequence of likelihood-based evaluations, but they are ill-suited for evaluation without a reference point, e.g., another model or an ideal score.

Without a discussion of the ideal score for a measure—which the creators of a measure are best suited to provide—users of the measure are left with an insufficient basis to determine if it is desirable to act on outcomes of the measure, and are thus inhibited from acting.

Unstated intended use. Almost half of all papers (47%) in our sample do not mention any intended use of their measure. Of the remaining papers, there are also cases where the intended use is only discussed in terms of future work that may be enabled by the paper, e.g.,

> "Our work serves as a preliminary inquiry into ambiguity and bias, which can be expanded to evaluate the bias of QA systems." – Mao et al. (2021)

A small subset of measures—from 34 papers—are more concrete and mention constraints that scope the use of the measure, by stating that their measure is to be used with one task or domain, e.g.,

> "We propose new methods to evaluate and mitigate gender bias *for languages with grammatical gender and bilingual word embeddings* [...]" – Zhou et al. (2019)

By providing this information, potential users of the measure can more easily determine whether it suits their use case.

Missing discussion around reliability. Surprisingly, we found that only 28 of the papers discuss *any* aspect of reliability, implicitly—i.e., by providing interval ranges or significance scores without accompanying discussion—or explicitly. Some work also uses reliability as a motivation to introduce a new measure (e.g., Nadeem et al., 2021; Alnegheimish et al., 2022; Kwon and Mihindukulasooriya, 2022; Pikuliak et al., 2023)), for example, by focusing on measures' robustness:

> "In this paper, we conduct an empirical study to *investigate the robustness* of the log-likelihood-based bias measure by paraphrasing the test sentences as in Figure 1 and analysing *if they produce consistent results*." – Kwon and Mihindukulasooriya (2022)

However, motivations around building more reliable measures do not necessarily translate into actually *studying* it. Only 42% of the papers that use reliability as a motivation for their work study the reliability of their methods. See Table 1 for full details.

Motivation	$\mathbf{R}_Y$	$\mathbf{R}_N$
Lack of reliability of existing measures	8	11
Measuring a missing or new bias	8	6
Measuring in a new setting or modality	14	16
Adjusting existing measures[11]	10	10
Measuring in a new language	12	15
No or unclear motivation	7	26
Total	59	84

Table 1: **Motivations provided for new measures.** Absolute counts in our collection (n=146) split into whether the authors discuss reliability (R_Y) or not (R_N).

6 Discussion

A considerable number of papers fail to provide crucial details about what motivates a bias measure and how it should be used. Therefore, we offer several suggestions for the development and dissemination of new bias measures.

▸ **Be clear about motivations, intended uses, and bias constructs.** Why is a new bias measure needed? How does it differ from existing measures, and which issue(s) does it address? What is the bias construct being operationalized? Without explicitly answering such questions, it is impossible to assess whether a measure addresses the need it is implicitly aimed at, or to which use cases it is well-suited.

Indeed, any proposed measure is accompanied, implicitly or explicitly, with an intended use; most papers introduce measures in the context of their use for some model or system. What many papers leave unstated is to which other settings a measure may be applied—e.g., other models, domains, or (social) contexts, if any. Therefore, practitioners are, more often than not, unable to assess a measure's suitability for their use cases.

Similarly, providing explicit reasoning about the the construct a measure is intended to capture can help prevent conflation between the conceptualization and the operationalization of a construct (e.g., Jacobs and Wallach, 2021; Blodgett et al., 2021). We argue that clearly articulating the underlying construct can additionally help in defining the measure's intended scope and use. This is particularly important given that recent work has shown that bias measures are often so closely tied to specific use cases that they cannot be reused across tasks, datasets, or languages (Delobelle et al., 2022; Or-

[11] By "adjusting" we understand measures which were created by modifying existing measures, e.g., using a different statistical test.

gad and Belinkov, 2022).

Lack of clarity in motivation, intended use, or bias construct may lead practitioners to adopt measures that are poorly matched to their own use cases (e.g., without realizing that some of the design choices for the measure are tied to particular intended uses), or forego measures that would have been appropriate due to insufficient knowledge of their applicability.

▸ **Relate measures' results with impacts or harms arising from models or systems.** Most papers in our sample report either an interval of variation, an ideal result, or both. Values for an interval or ideal result represent some subjective assessment of the desirability of model or system behaviors at those values (Waseem et al., 2021). However, we find that most papers only implicitly relate measures' results and model or system behaviors. We therefore encourage the creators of measures to ground the values their measure can take—at least for the ideal result and extrema—in the expected behaviors, and resulting impacts or harms, that a model or system might produce. Such information can help future users determine the appropriateness of a measure for their purposes, and how to act in cases of deviation from ideal results or relative distances from the extrema.

▸ **Always assess reliability.** Only a very small number of papers presenting bias measures formally assess their reliability. Although this issue has been raised before (e.g., Delobelle et al., 2022; Orgad et al., 2022), it remains concerning. Measurement processes that provide a basis for informed decision-making by their nature rely on reproducibility and predictable variation in measures' results for their external justification. The lack of information on reliability may ultimately lead users of a measure to conclude that they cannot act on it due to a lack of trust in the outcome.

▸ **Consider the target audience.** When developing a bias measure, it is important to also consider the stakeholders that might be using the measure, and which actions are afforded to each stakeholder. For example, while unbounded measures can be useful for system developers, they may not be very useful to regulators, decision-makers within a company, or individuals potentially using a system if they are not grounded in actual impacts or harms.

Moreover, although stakeholders may be—in principle—equally able to take an action on the basis of some bias measures, which actions they are afforded may differ based on their ability to directly intervene in the system. For instance, individuals' actions may be limited to refusal (Gangadharan, 2021) and collective action as means for changing a system, while developers, companies, and regulators can engage in more direct processes. Developers can address biases in models and systems; companies can allocate resources for addressing them, delay deployment, or retire models and systems entirely; and regulators can engage in regulatory processes to develop new regulation or apply existing regulation. When developing a bias measure, it may therefore also be appropriate to consider which stakeholder(s) the measure should enable to take action.

7 Conclusion

We introduce actionability of bias measures, identify several desiderata for the actionability, and annotate 146 papers in the NLP literature for these desiderata, finding that much information required for actionability is under-specified or unstated. This finding suggests that current measures may not enable practitioners to meaningfully act on their results. We provide recommendations for future work that we hope can support the development of actionable bias measures, and believe that our desiderata can serve as a starting point for broader discussions on how we assess bias in models and systems, and more broadly help minimize the disparity between research artifacts and their practical uptake. Moreover, although bias measures have been the focal point of our intervention, further work could explore how our framework might extend to other measurement instruments. Such measurement instruments may facilitate different possible actions or interventions than bias measures, and actionability for those instruments may demand different desiderata.

Limitations

This paper comes with several limitations.

Perspectives. When selecting our desiderata, we reviewed literature within NLP and related fields, which could limit the breadth of the desiderata for actionability that we identified. Therefore, it is possible that we overlooked potential desiderata for actionable bias measures or provided a definition of actionability that is too loose or too stringent Moreover, depending on the context of use, our desiderata might be "necessary" but not "sufficient"—i.e., even if all desiderata are met,

measures' results might still not provide actionable insights. We view this work as another point in a longer discourse on the conceptual and practical lack of clarity around bias measures (see also Blodgett et al., 2020, 2021; Jacobs and Wallach, 2021).

Methods. Our procedure of sampling papers from the ACL Anthology has inherent limitations. Although we incorporate some papers from other sources, as discussed in Section 4, we primarily focus on the ACL community, which prevents the inclusion of significant contributions and perspectives from machine learning venues. However, our primary objective was to examine how authors discuss bias measures in the NLP literature, what information they choose to present, and whether this information is sufficient for taking informed action on the basis of the outcome of a measure; thus, we conducted a large scale analysis of 146 papers in the ACL Anthology prior to June 2024, ensuring that our analysis is appropriate for language technologies.

Ethics statement

Our paper assumes that language technologies will be deployed in contexts where they are applied to human data and may produce socially discriminatory outcomes. We further assume that there exist some individuals or organizations that would be interested taking meaningful steps to measure and mitigate the production of (algorithmic) discrimination. Under such assumptions, providing mechanisms and processes for determining the degree to which a measure is actionable can be one factor in choosing bias measures to apply. Moreover, measures with high degrees of actionability can help facilitate trust in models and systems that are deployed. Finally, due to our methods' limitations and our own subjectivities, the desiderata and recommendations that we provide should be treated as a starting point, rather than as conclusive.

Acknowledgments

We thank the reviewers, the members of the SARDINE, DTAI, HU Berlin's ML and MilaNLP research groups, and Sonja Mei Wang for their insightful comments. Giuseppe Attanasio was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI) and by Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020. He conducted part of the work as a member of the MilaNLP group at Bocconi University, Milan. Pieter Delobelle received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme. He is supported by the Research Foundation - Flanders (FWO) under EOS No. 30992574 (VeriLearn) and received a grant from "Interne Fondsen KU Leuven/Internal Funds KU Leuven." He conducted part of the work as a visitor to the MilaNLP group at Bocconi University, Milan. Debora Nozza was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 101116095, PERSONAE). Debora Nozza is a member of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis (BIDSA).

References

Tosin Adewumi, Isabella Södergren, Lama Alkhaled, Sana Al-azzawi, Foteini Simistira Liwicki, and Marcus Liwicki. 2023. Bipol: Multi-axes evaluation of bias with explainability in benchmark datasets. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1–10, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Arshiya Aggarwal, Jiao Sun, and Nanyun Peng. 2022. Towards robust NLG bias evaluation with syntactically-diverse prompts. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6022–6032, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jaimeen Ahn and Alice Oh. 2021. Mitigating language-dependent ethnic bias in BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 533–549, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ifeoma Ajunwa, Sorelle Friedler, Carlos E Scheidegger, and Suresh Venkatasubramanian. 2016. Hiring by algorithm: predicting and preventing disparate impact. *Available at SSRN*, 2746078:29.

Sarah Alnegheimish, Alicia Guo, and Yi Sun. 2022. Using natural sentence prompts for understanding biases in language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2824–2830, Seattle, United States. Association for Computational Linguistics.

Anonymous. 2016. Incident number 37. *AI Incident Database*.

Maria Antoniak and David Mimno. 2021. Bad seeds: Evaluating lexical methods for bias measurement. 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 1889–1904, Online. Association for Computational Linguistics.

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based attention regularization frees unintended bias mitigation from lists. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1105–1119, Dublin, Ireland. Association for Computational Linguistics.

Giuseppe Attanasio, Flor Miriam Plaza del Arco, Debora Nozza, and Anne Lauscher. 2023. A tale of pronouns: Interpretability informs gender bias mitigation for fairer instruction-tuned machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3996–4014, Singapore. Association for Computational Linguistics.

Senthil Kumar B, Pranav Tiwari, Aman Chandra Kumar, and Aravindan Chandrabose. 2022. Casteism in India, but not racism - a study of bias in word embeddings of Indian languages. In *Proceedings of the First Workshop on Language Technology and Resources for a Fair, Inclusive, and Safe Society within the 13th Language Resources and Evaluation Conference*, pages 1–7, Marseille, France. European Language Resources Association.

Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. 2022. How well can text-to-image generative models understand ethical natural language interventions? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1370, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. 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 1941–1955, Online. Association for Computational Linguistics.

Ainhize Barrainkua, Paula Gordaliza, Jose A Lozano, and Novi Quadrianto. 2023. Preserving the fairness guarantees of classifiers in changing environments: a survey. *ACM Computing Surveys*.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy. Association for Computational Linguistics.

Christine Basta, Marta R Costa-Jussa, and Noe Casas. 2021. Extensive study on the underlying gender bias in contextualized word embeddings. *Neural Computing and Applications*, 33(8):3371–3384.

Somnath Basu Roy Chowdhury, Sayan Ghosh, Yiyuan Li, Junier Oliva, Shashank Srivastava, and Snigdha Chaturvedi. 2021. Adversarial scrubbing of demographic information for text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 550–562, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lisa Bauer, Ninareh Mehrabi, Palash Goyal, Kai-Wei Chang, Aram Galstyan, and Rahul Gupta. 2024. Believe: Belief-enhanced instruction generation and augmentation for zero-shot bias mitigation. In *NAACL 2024 Workshop on TrustNLP*.

Hugo Berg, Siobhan Hall, Yash Bhalgat, Hannah Kirk, Aleksandar Shtedritski, and Max Bain. 2022. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 806–822, Online only. Association for Computational Linguistics.

Jayadev Bhaskaran and Isha Bhallamudi. 2019. Good secretaries, bad truck drivers? occupational gender stereotypes in sentiment analysis. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 62–68, Florence, Italy. Association for Computational Linguistics.

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Re-contextualizing fairness in NLP: The case of India. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 727–740, Online only. Association for Computational Linguistics.

El Moatez Billah Nagoudi, Muhammad Abdul-Mageed, AbdelRahim Elmadany, Alcides Inciarte, and Md Tawkat Islam Khondaker. 2023. JASMINE: Arabic GPT models for few-shot learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16721–16744, Singapore. Association for Computational Linguistics.

Abeba Birhane, Ryan Steed, Victor Ojewale, Briana Vecchione, and Inioluwa Deborah Raji. 2024. AI

auditing: The Broken Bus on the Road to AI Accountability. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 612–643. IEEE Computer Society.

Emily Black, Talia Gillis, and Zara Yasmine Hall. 2024. D-hacking. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 602–615, New York, NY, USA. Association for Computing Machinery.

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

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

Conrad Borchers, Dalia Gala, Benjamin Gilburt, Eduard Oravkin, Wilfried Bounsi, Yuki M Asano, and Hannah Kirk. 2022. Looking for a handsome carpenter! debiasing GPT-3 job advertisements. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 212–224, Seattle, Washington. Association for Computational Linguistics.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Laura Cabello, Emanuele Bugliarello, Stephanie Brandl, and Desmond Elliott. 2023. Evaluating bias and fairness in gender-neutral pretrained vision-and-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8465–8483, Singapore. Association for Computational Linguistics.

Laura Cabello Piqueras and Anders Søgaard. 2022. Are pretrained multilingual models equally fair across languages? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3597–3605, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

António Câmara, Nina Taneja, Tamjeed Azad, Emily Allaway, and Richard Zemel. 2022. Mapping the multilingual margins: Intersectional biases of sentiment analysis systems in English, Spanish, and Arabic. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 90–106, Dublin, Ireland. Association for Computational Linguistics.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Schwemer, and Anders Søgaard. 2022. FairLex: A multilingual benchmark for evaluating fairness in legal text processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4389–4406, Dublin, Ireland. Association for Computational Linguistics.

Rodrigo Alejandro Chávez Mulsa and Gerasimos Spanakis. 2020. Evaluating bias in Dutch word embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 56–71, Barcelona, Spain (Online). Association for Computational Linguistics.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. CoMPosT: Characterizing and evaluating caricature in LLM simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10853–10875, Singapore. Association for Computational Linguistics.

Chloe Ciora, Nur Iren, and Malihe Alikhani. 2021. Examining covert gender bias: A case study in Turkish and English machine translation models. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 55–63, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Marta Costa-jussà, Pierre Andrews, Eric Smith, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Daniel Licht, and Carleigh Wood. 2023. Multilingual holistic bias: Extending descriptors and patterns to unveil demographic biases in languages at scale. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14141–14156, Singapore. Association for Computational Linguistics.

Marta R. Costa-jussà, Christine Basta, and Gerard I. Gállego. 2022. Evaluating gender bias in speech translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2141–2147, Marseille, France. European Language Resources Association.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Mayukh Das and Wolf Tilo Balke. 2022. Quantifying bias from decoding techniques in natural language generation. In *Proceedings of the 29th International Conference on Computational Linguistics*,

pages 1311–1323, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jeffrey Dastin. 2018. Insight - Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*.

Hillary Dawkins. 2021. Second order WinoBias (SoWinoBias) test set for latent gender bias detection in coreference resolution. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 103–111, Online. Association for Computational Linguistics.

Erenay Dayanik, Ngoc Thang Vu, and Sebastian Padó. 2022. Bias identification and attribution in NLP models with regression and effect sizes. In *Northern European Journal of Language Technology, Volume 8*, Copenhagen, Denmark. Northern European Association of Language Technology.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2232–2242, Online. Association for Computational Linguistics.

Nicholas Deas, Jessica Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of African American language bias in natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6805–6824, Singapore. Association for Computational Linguistics.

Pieter Delobelle and Bettina Berendt. 2023. FairDistillation: Mitigating Stereotyping in Language Models. In *Machine Learning and Knowledge Discovery in Databases*, volume 13714, pages 638–654, Cham. Springer International Publishing. Series Title: Lecture Notes in Computer Science.

Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1693–1706, Seattle, United States. Association for Computational Linguistics.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270, Singapore. Association for Computational Linguistics.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. 2021. OSCaR: Orthogonal subspace correction and rectification of biases in word embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5034–5050, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. On measures of biases and harms in NLP. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 246–267, Online only. Association for Computational Linguistics.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020a. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020b. Multi-dimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online. Association for Computational Linguistics.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv: Machine Learning*.

Yupei Du, Yuanbin Wu, and Man Lan. 2019. Exploring human gender stereotypes with word association test. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6133–6143, Hong Kong, China. Association for Computational Linguistics.

Fatma Elsafoury, Steve R. Wilson, Stamos Katsigiannis, and Naeem Ramzan. 2022. SOS: Systematic offensive stereotyping bias in word embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1263–1274, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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

David Esiobu, Xiaoqing Tan, Saghar Hosseini, Megan Ung, Yuchen Zhang, Jude Fernandes, Jane Dwivedi-Yu, Eleonora Presani, Adina Williams, and Eric Smith. 2023. ROBBIE: Robust bias evaluation of large generative language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3764–3814, Singapore. Association for Computational Linguistics.

Cristina España-Bonet and Alberto Barrón-Cedeño. 2022. The (undesired) attenuation of human biases by multilinguality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2056–2077, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics.

Susan T Fiske. 2018. Stereotype content: Warmth and competence endure. *Current directions in psychological science*, 27(2):67–73.

Scott Friedman, Sonja Schmer-Galunder, Anthony Chen, and Jeffrey Rye. 2019. Relating word embedding gender biases to gender gaps: A cross-cultural analysis. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 18–24, Florence, Italy. Association for Computational Linguistics.

Marco Gaido, Beatrice Savoldi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2020. Breeding gender-aware direct speech translation systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3951–3964, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Seeta Gangadharan. 2021. *4. Digital Exclusion: A Politics of Refusal*, pages 113–140. University of Chicago Press, Chicago.

Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, Florence, Italy. Association for Computational Linguistics.

Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2020. Towards understanding gender bias in relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2943–2953, Online. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Adam Lopez, Roi Blanco, and Diego Marcheggiani. 2023a. Bias beyond English: Counterfactual tests for bias in sentiment analysis in four languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4458–4468, Toronto, Canada. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. 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 1926–1940, Online. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Björn Ross, and Adam Lopez. 2023b. Cross-lingual transfer can worsen bias in sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5691–5704, Singapore. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Eddie Ungless, Esma Balkir, and Su Lin Blodgett. 2023c. This prompt is measuring <mask>: evaluating bias evaluation in language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2209–2225, Toronto, Canada. Association for Computational Linguistics.

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 American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Steven N. Goodman, Daniele Fanelli, and John P. A. Ioannidis. 2016. What does research reproducibility mean? *Science Translational Medicine*, 8(341).

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 122–133, New York, NY, USA. Association for Computing Machinery.

Rishav Hada, Agrima Seth, Harshita Diddee, and Kalika Bali. 2023. "fifty shades of bias": Normative ratings of gender bias in GPT generated English text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1862–1876, Singapore. Association for Computational Linguistics.

Oussama Hansal, Ngoc Tan Le, and Fatiha Sadat. 2022. Indigenous language revitalization and the dilemma of gender bias. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 244–254, Seattle, Washington. Association for Computational Linguistics.

Saga Hansson, Konstantinos Mavromatakis, Yvonne Adesam, Gerlof Bouma, and Dana Dannélls. 2021. The Swedish Winogender dataset. In *Proceedings*

of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), pages 452–459, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Christian Hardmeier, Marta R. Costa-jussà, Kellie Webster, Will Radford, and Su Lin Blodgett. 2021. How to Write a Bias Statement: Recommendations for Submissions to the Workshop on Gender Bias in NLP. ArXiv:2104.03026 [cs].

Samhita Honnavalli, Aesha Parekh, Lily Ou, Sophie Groenwold, Sharon Levy, Vicente Ordonez, and William Yang Wang. 2022. Towards understanding gender-seniority compound bias in natural language generation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1665–1670, Marseille, France. European Language Resources Association.

Saghar Hosseini, Hamid Palangi, and Ahmed Hassan Awadallah. 2023. An empirical study of metrics to measure representational harms in pre-trained language models. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 121–134, Toronto, Canada. Association for Computational Linguistics.

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J. Paul. 2020. Multilingual Twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1440–1448, Marseille, France. European Language Resources Association.

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

Sophie Jentzsch and Cigdem Turan. 2022. Gender bias in BERT - measuring and analysing biases through sentiment rating in a realistic downstream classification task. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 184–199, Seattle, Washington. Association for Computational Linguistics.

Sullam Jeoung, Yubin Ge, and Jana Diesner. 2023. StereoMap: Quantifying the awareness of human-like stereotypes in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12236–12256, Singapore. Association for Computational Linguistics.

Akshita Jha, Aida Mostafazadeh Davani, Chandan K Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. SeeGULL: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9851–9870, Toronto, Canada. Association for Computational Linguistics.

Meichun Jiao and Ziyang Luo. 2021. Gender bias hidden behind Chinese word embeddings: The case of Chinese adjectives. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 8–15, Online. Association for Computational Linguistics.

Anna Jørgensen and Anders Søgaard. 2021. Evaluation of summarization systems across gender, age, and race. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 51–56, Online and in Dominican Republic. Association for Computational Linguistics.

Kenneth Joseph and Jonathan Morgan. 2020. When do word embeddings accurately reflect surveys on our beliefs about people? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4392–4415, Online. Association for Computational Linguistics.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022. Gender bias in masked language models for multiple languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2740–2750, Seattle, United States. Association for Computational Linguistics.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *International Conference on Machine Learning*, pages 2564–2572. PMLR.

Simran Khanuja, Sebastian Ruder, and Partha Talukdar. 2023. Evaluating the diversity, equity, and inclusion of NLP technology: A case study for Indian languages. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1763–1777, Dubrovnik, Croatia. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020. Gender coreference and bias evaluation at WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 357–364, Online. Association for Computational Linguistics.

Abdullatif Köksal, Omer Yalcin, Ahmet Akbiyik, M. Kilavuz, Anna Korhonen, and Hinrich Schuetze. 2023. Language-agnostic bias detection in language models with bias probing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12735–12747, Singapore. Association for Computational Linguistics.

Daan Kolkman. 2020. F** k the algorithm?: what the world can learn from the uk's a-level grading fiasco. *Impact of Social Sciences Blog*.

Satyapriya Krishna, Rahul Gupta, Apurv Verma, Jwala Dhamala, Yada Pruksachatkun, and Kai-Wei Chang. 2022. Measuring fairness of text classifiers via prediction sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5830–5842, Dublin, Ireland. Association for Computational Linguistics.

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

Alexander Kwako, Yixin Wan, Jieyu Zhao, Kai-Wei Chang, Li Cai, and Mark Hansen. 2022. Using item response theory to measure gender and racial bias of a BERT-based automated English speech assessment system. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 1–7, Seattle, Washington. Association for Computational Linguistics.

Bum Chul Kwon and Nandana Mihindukulasooriya. 2022. An empirical study on pseudo-log-likelihood bias measures for masked language models using paraphrased sentences. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pages 74–79, Seattle, U.S.A. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen McKeown, and Tatsunori Hashimoto. 2023. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3206–3219, Dubrovnik, Croatia. Association for Computational Linguistics.

Stefan Larsson and Fredrik Heintz. 2020. Transparency in artificial intelligence. *Internet Policy Review*, 9(2).

Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Minwoo Lee, Hyukhun Koh, Kang-il Lee, Dongdong Zhang, Minsung Kim, and Kyomin Jung. 2023. Target-agnostic gender-aware contrastive learning for mitigating bias in multilingual machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16825–16839, Singapore. Association for Computational Linguistics.

Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2470–2480, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. UNQOVERing stereotyping biases via underspecified questions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489, Online. Association for Computational Linguistics.

Yizhi Li, Ge Zhang, Bohao Yang, Chenghua Lin, Anton Ragni, Shi Wang, and Jie Fu. 2022. HERB: Measuring hierarchical regional bias in pre-trained language models. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 334–346, Online only. Association for Computational Linguistics.

Q. Vera Liao and Jennifer Wortman Vaughan. 2024. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *Harvard Data Science Review*, (Special Issue 5).

Inna Lin, Lucille Njoo, Anjalie Field, Ashish Sharma, Katharina Reinecke, Tim Althoff, and Yulia Tsvetkov. 2022. Gendered mental health stigma in masked language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2152–2170, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2020. Does gender matter? towards fairness in dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4403–4416, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yu Lu Liu, Meng Cao, Su Lin Blodgett, Jackie Chi Kit Cheung, Alexandra Olteanu, and Adam Trischler. 2023. Responsible AI considerations in text summarization research: A review of current practices. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6246–6261, Singapore. Association for Computational Linguistics.

Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, Thomas Wang, Nouamane Tazi, Teven Scao, Thomas Wolf, Osma Suominen, Samuli Sairanen, Mikko Merioksa, Jyrki Heinonen, Aija Vahtola, Samuel Antao, and Sampo Pyysalo. 2023. FinGPT: Large generative models for a small language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2710–2726, Singapore. Association for Computational Linguistics.

Weicheng Ma, Brian Chiang, Tong Wu, Lili Wang, and Soroush Vosoughi. 2023. Intersectional stereotypes in large language models: Dataset and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8589–8597, Singapore. Association for Computational Linguistics.

Gaurav Maheshwari, Aurélien Bellet, Pascal Denis, and Mikaela Keller. 2023. Fair without leveling down: A new intersectional fairness definition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9018–9032, Singapore. Association for Computational Linguistics.

Manuj Malik and Richard Johansson. 2022. Controlling for stereotypes in multimodal language model evaluation. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 263–271, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. Socially aware bias measurements for Hindi language representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1041–1052, Seattle, United States. Association for Computational Linguistics.

Marta Marchiori Manerba and Sara Tonelli. 2021. Fine-grained fairness analysis of abusive language detection systems with CheckList. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 81–91, Online. Association for Computational Linguistics.

Courtney Mansfield, Amandalynne Paullada, and Kristen Howell. 2022. Behind the mask: Demographic bias in name detection for PII masking. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 76–89, Dublin, Ireland. Association for Computational Linguistics.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Andrew Mao, Naveen Raman, Matthew Shu, Eric Li, Franklin Yang, and Jordan Boyd-Graber. 2021. Eliciting bias in question answering models through ambiguity. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 92–99, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sandra Martinková, Karolina Stanczak, and Isabelle Augenstein. 2023. Measuring gender bias in West Slavic language models. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 146–154, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, 8(1):141–163.

Aida Mostafazadeh Davani, Ali Omrani, Brendan Kennedy, Mohammad Atari, Xiang Ren, and Morteza Dehghani. 2021. Improving counterfactual generation for fair hate speech detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 92–101, Online. Association for Computational Linguistics.

Anjishnu Mukherjee, Chahat Raj, Ziwei Zhu, and Antonios Anastasopoulos. 2023. Global Voices, local biases: Socio-cultural prejudices across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15828–15845, Singapore. Association for Computational Linguistics.

Robert Munro and Alex (Carmen) Morrison. 2020. Detecting independent pronoun bias with partially-synthetic data generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2011–2017, Online. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. 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 5356–5371, Online. Association for Computational Linguistics.

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

Isar Nejadgholi, Esma Balkir, Kathleen Fraser, and Svetlana Kiritchenko. 2022. Towards procedural fairness: Uncovering biases in how a toxic language classifier

uses sentiment information. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 225–237, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531, Dublin, Ireland. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, Anne Lauscher, and Dirk Hovy. 2022. Measuring harmful sentence completion in language models for LGBTQIA+ individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 26–34, Dublin, Ireland. Association for Computational Linguistics.

Dario Onorati, Elena Sofia Ruzzetti, Davide Venditti, Leonardo Ranaldi, and Fabio Massimo Zanzotto. 2023. Measuring bias in instruction-following models with P-AT. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8006–8034, Singapore. Association for Computational Linguistics.

Hadas Orgad and Yonatan Belinkov. 2022. Choose your lenses: Flaws in gender bias evaluation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 151–167, Seattle, Washington. Association for Computational Linguistics.

Hadas Orgad, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. 2022. How gender debiasing affects internal model representations, and why it matters. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2602–2628, Seattle, United States. Association for Computational Linguistics.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. 2021. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Andrea Piergentili, Beatrice Savoldi, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2023. Hi guys or hi folks? benchmarking gender-neutral machine translation with the GeNTE corpus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14124–14140, Singapore. Association for Computational Linguistics.

Matúš Pikuliak, Ivana Beňová, and Viktor Bachratý. 2023. In-depth look at word filling societal bias measures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3648–3665, Dubrovnik, Croatia. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5740–5745, Hong Kong, China. Association for Computational Linguistics.

Nirmalendu Prakash and Roy Ka-Wei Lee. 2023. Layered bias: Interpreting bias in pretrained large language models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 284–295, Singapore. Association for Computational Linguistics.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Michael Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9496–9521, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 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*, pages 33–44, Barcelona Spain. ACM.

Krithika Ramesh, Gauri Gupta, and Sanjay Singh. 2021. Evaluating gender bias in Hindi-English machine translation. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 16–23, Online. Association for Computational Linguistics.

Haocong Rao, Cyril Leung, and Chunyan Miao. 2023. Can ChatGPT assess human personalities? a general evaluation framework. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1184–1194, Singapore. Association for Computational Linguistics.

Adithya Renduchintala, Denise Diaz, Kenneth Heafield, Xian Li, and Mona Diab. 2021. Gender bias amplification during speed-quality optimization in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 99–109, Online. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Candace Ross, Boris Katz, and Andrei Barbu. 2021. Measuring social biases in grounded vision and language embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 998–1008, Online. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Gabriele Ruggeri and Debora Nozza. 2023. A multidimensional study on bias in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6445–6455, Toronto, Canada. Association for Computational Linguistics.

Ahmed Sabir and Lluís Padró. 2023. Women wearing lipstick: Measuring the bias between an object and its related gender. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4234–4240, Singapore. Association for Computational Linguistics.

Sandra Sandoval, Jieyu Zhao, Marine Carpuat, and Hal Daumé III. 2023. A rose by any other name would not smell as sweet: Social bias in names mistranslation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3933–3945, Singapore. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2022. Under the morphosyntactic lens: A multifaceted evaluation of gender bias in speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1807–1824, Dublin, Ireland. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2023. Test suites task: Evaluation of gender fairness in MT with MuST-SHE and INES. In *Proceedings of the Eighth Conference on Machine Translation*, pages 252–262, Singapore. Association for Computational Linguistics.

Nikil Selvam, Sunipa Dev, Daniel Khashabi, Tushar Khot, and Kai-Wei Chang. 2023. The tail wagging the dog: Dataset construction biases of social bias benchmarks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1373–1386, Toronto, Canada. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Pushpdeep Singh. 2023. Gender inflected or bias inflicted: On using grammatical gender cues for bias evaluation in machine translation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 17–23, Nusa Dua, Bali. Association for Computational Linguistics.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nasim Sobhani, Kinshuk Sengupta, and Sarah Jane Delany. 2023. Measuring gender bias in natural language processing: Incorporating gender-neutral linguistic forms for non-binary gender identities in abusive speech detection. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1121–1131, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Irene Solaiman. 2023. The gradient of generative ai release: Methods and considerations. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*.

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

Victor Steinborn, Philipp Dufter, Haris Jabbar, and Hinrich Schuetze. 2022. An information-theoretic approach and dataset for probing gender stereotypes in multilingual masked language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 921–932, Seattle, United States. Association for Computational Linguistics.

Xavier Suau, Pieter Delobelle, Katherine Metcalf, Armand Joulin, Nicholas Apostoloff, Luca Zappella, and Pau Rodriguez. 2024. Whispering experts: Neural interventions for toxicity mitigation in language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 46843–46867. PMLR.

Tianxiang Sun, Junliang He, Xipeng Qiu, and Xuanjing Huang. 2022. BERTScore is unfair: On social bias in language model-based metrics for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3726–3739, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chris Sweeney and Maryam Najafian. 2019. A transparent framework for evaluating unintended demographic bias in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1662–1667, Florence, Italy. Association for Computational Linguistics.

Karolina Sztandar-Sztanderska and Marianna Zieleńska. 2022. When a Human Says "No" to a Computer: Frontline Oversight of the Profiling Algorithm in Public Employment Services in Poland. *Sozialer Fortschritt*, 71(6-7):465–487.

Zeerak Talat and Anne Lauscher. 2022. Back to the Future: On Potential Histories in NLP. *Preprint*, arXiv:2210.06245.

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

Yi Chern Tan and L. Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Ewoenam Kwaku Tokpo, Pieter Delobelle, Bettina Berendt, and Toon Calders. 2023. How far can it go? on intrinsic gender bias mitigation for text classification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3418–3433, Dubrovnik, Croatia. Association for Computational Linguistics.

Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2023. Measuring normative and descriptive biases in language models using census data. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2242–2248, Dubrovnik, Croatia. Association for Computational Linguistics.

Bertille Triboulet and Pierrette Bouillon. 2023. Evaluating the impact of stereotypes and language combinations on gender bias occurrence in NMT generic systems. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 62–70, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Jonas-Dario Troles and Ute Schmid. 2021. Extending challenge sets to uncover gender bias in machine translation: Impact of stereotypical verbs and adjectives. In *Proceedings of the Sixth Conference on Machine Translation*, pages 531–541, Online. Association for Computational Linguistics.

Francisco Valentini, Germán Rosati, Diego Fernandez Slezak, and Edgar Altszyler. 2022. The undesirable dependence on frequency of gender bias metrics based on word embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5086–5092, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Francielle Vargas, Isabelle Carvalho, Ali Hürriyetoğlu, Thiago Pardo, and Fabrício Benevenuto. 2023. Socially responsible hate speech detection: Can classifiers reflect social stereotypes? In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1187–1196, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Eric Peter Wairagala, Jonathan Mukiibi, Jeremy Francis Tusubira, Claire Babirye, Joyce Nakatumba-Nabende, Andrew Katumba, and Ivan Ssenkungu. 2022. Gender bias evaluation in Luganda-English

machine translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 274–286, Orlando, USA. Association for Machine Translation in the Americas.

Thiemo Wambsganss, Xiaotian Su, Vinitra Swamy, Seyed Neshaei, Roman Rietsche, and Tanja Käser. 2023. Unraveling downstream gender bias from large language models: A study on AI educational writing assistance. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10275–10288, Singapore. Association for Computational Linguistics.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023a. "kelly is a warm person, joseph is a role model": Gender biases in LLM-generated reference letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748, Singapore. Association for Computational Linguistics.

Yixin Wan, Jieyu Zhao, Aman Chadha, Nanyun Peng, and Kai-Wei Chang. 2023b. Are personalized stochastic parrots more dangerous? evaluating persona biases in dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9677–9705, Singapore. Association for Computational Linguistics.

Jialu Wang, Xinyue Liu, Zonglin Di, Yang Liu, and Xin Wang. 2023a. T2IAT: Measuring valence and stereotypical biases in text-to-image generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2560–2574, Toronto, Canada. Association for Computational Linguistics.

Jun Wang, Benjamin Rubinstein, and Trevor Cohn. 2022. Measuring and mitigating name biases in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2576–2590, Dublin, Ireland. Association for Computational Linguistics.

Nan Wang, Qifan Wang, Yi-Chia Wang, Maziar Sanjabi, Jingzhou Liu, Hamed Firooz, Hongning Wang, and Shaoliang Nie. 2023b. COFFEE: Counterfactual fairness for personalized text generation in explainable recommendation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13258–13275, Singapore. Association for Computational Linguistics.

Zeerak Waseem, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. 2021. Disembodied Machine Learning: On the Illusion of Objectivity in NLP. *arXiv preprint*. Number: arXiv:2101.11974 arXiv:2101.11974 [cs].

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2021. Measuring and Reducing Gendered Correlations in Pre-trained Models. ArXiv:2010.06032 [cs].

Zhongbin Xie, Vid Kocijan, Thomas Lukasiewicz, and Oana-Maria Camburu. 2023. Counter-GAP: Counterfactual bias evaluation through gendered ambiguous pronouns. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3761–3773, Dubrovnik, Croatia. Association for Computational Linguistics.

Kai-Ching Yeh, Jou-An Chi, Da-Chen Lian, and Shu-Kai Hsieh. 2023. Evaluating interfaced LLM bias. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 292–299, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Catherine Yeo and Alyssa Chen. 2020. Defining and Evaluating Fair Natural Language Generation. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 107–109, Seattle, USA. Association for Computational Linguistics.

Mahdi Zakizadeh, Kaveh Miandoab, and Mohammad Pilehvar. 2023. DiFair: A benchmark for disentangled assessment of gender knowledge and bias. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1897–1914, Singapore. Association for Computational Linguistics.

Jieyu Zhao and Kai-Wei Chang. 2020. LOGAN: Local group bias detection by clustering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1968–1977, Online. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Jingyan Zhou, Jiawen Deng, Fei Mi, Yitong Li, Yasheng Wang, Minlie Huang, Xin Jiang, Qun Liu, and Helen Meng. 2022a. Towards identifying social bias in dialog systems: Framework, dataset, and benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3576–3591, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. 2022b. Deconstructing NLG evaluation: Evaluation practices, assumptions, and their implications. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 314–324, Seattle, United States. Association for Computational Linguistics.

Kankan Zhou, Eason Lai, and Jing Jiang. 2022c. VLStereoSet: A study of stereotypical bias in pre-trained vision-language models. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 527–538, Online only. Association for Computational Linguistics.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5276–5284, Hong Kong, China. Association for Computational Linguistics.

Yi Zhou, Masahiro Kaneko, and Danushka Bollegala. 2022d. Sense embeddings are also biased – evaluating social biases in static and contextualised sense embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1924–1935, Dublin, Ireland. Association for Computational Linguistics.

A Annotation Details

A.1 Annotated papers

In this section, we list the annotated papers grouped by their motivation for introducing a new metric.

Lack of reliability of existing metrics. Du et al. (2019); Ethayarajh et al. (2019); Joseph and Morgan (2020); Sap et al. (2020); Webster et al. (2021); Zhao and Chang (2020); Manerba and Tonelli (2021); Mostafazadeh Davani et al. (2021); Nadeem et al. (2021); Kwon and Mihindukulasooriya (2022); Aggarwal et al. (2022); Alnegheimish et al. (2022); Das and Balke (2022); Dayanik et al. (2022); Nejadgholi et al. (2022); Sun et al. (2022); Zhou et al. (2022d); Pikuliak et al. (2023); Jeoung et al. (2023).

Measuring bias in a new setting or modality. Tatman (2017); Rudinger et al. (2018); Zhao et al. (2018); Escudé Font and Costa-jussà (2019); Sheng et al. (2019); Stanovsky et al. (2019); Dinan et al. (2020a); Gaido et al. (2020); Gaut et al. (2020); Liu et al. (2020); Yeo and Chen (2020); Barikeri et al. (2021); Jørgensen and Søgaard (2021); Renduchintala et al. (2021); Ross et al. (2021); Berg et al. (2022); Borchers et al. (2022); Costa-jussà et al. (2022); Kwako et al. (2022); Malik and Johansson (2022); Mansfield et al. (2022); Parrish et al. (2022); Zhou et al. (2022c); Cabello et al. (2023); Hosseini et al. (2023); Onorati et al. (2023); Ruggeri and Nozza (2023); Wan et al. (2023a,b); Wang et al. (2023b,a); Guo and Caliskan (2021).

Adjusting or improving an existing metric. May et al. (2019); Garimella et al. (2019); Kurita et al. (2019); Manzini et al. (2019); Dinan et al. (2020b); Munro and Morrison (2020); Basta et al. (2021); de Vassimon Manela et al. (2021); Levy et al. (2021); Troles and Schmid (2021); Qian et al. (2022); Valentini et al. (2022); Zhou et al. (2022a); Esiobu et al. (2023); Hada et al. (2023); Ma et al. (2023); Maheshwari et al. (2023); Prakash and Lee (2023); Xie et al. (2023); Zakizadeh et al. (2023).

Measuring a missing or new type of bias. Tan and Celis (2019); Ahn and Oh (2021); Dawkins (2021); Nozza et al. (2021); Elsafoury et al. (2022); Câmara et al. (2022); Honnavalli et al. (2022); Li et al. (2022); Lin et al. (2022); Nozza et al. (2022); Cheng et al. (2023); Goldfarb-Tarrant et al. (2023b); Piergentili et al. (2023); Sandoval et al. (2023); Savoldi et al. (2023); Sobhani et al. (2023).

Measuring bias in a new language. Zhou et al. (2019); Chávez Mulsa and Spanakis (2020); Huang et al. (2020); Kocmi et al. (2020); Hansson et al. (2021); Jiao and Luo (2021); Ramesh et al. (2021); Malik et al. (2022); B et al. (2022); Bhatt et al. (2022); Cabello Piqueras and Søgaard (2022); España-Bonet and Barrón-Cedeño (2022); Hansal et al. (2022); Kaneko et al. (2022); Névéol et al. (2022); Steinborn et al. (2022); Wairagala et al. (2022); Billah Nagoudi et al. (2023); Costa-jussà et al. (2023); Deas et al. (2023); Goldfarb-Tarrant et al. (2023a); Khanuja et al. (2023); Köksal et al. (2023); Martinková et al. (2023); Mukherjee et al. (2023); Singh (2023); Wambsganss et al. (2023).

Unclear or no motivation. Kiritchenko and Mohammad (2018); Basta et al. (2019); Bhaskaran and Bhallamudi (2019); Bordia and Bowman (2019); Friedman et al. (2019); Prabhakaran et al. (2019); Sweeney and Najafian (2019); Zhao et al. (2019); Bartl et al. (2020); Li et al. (2020); Nangia et al.

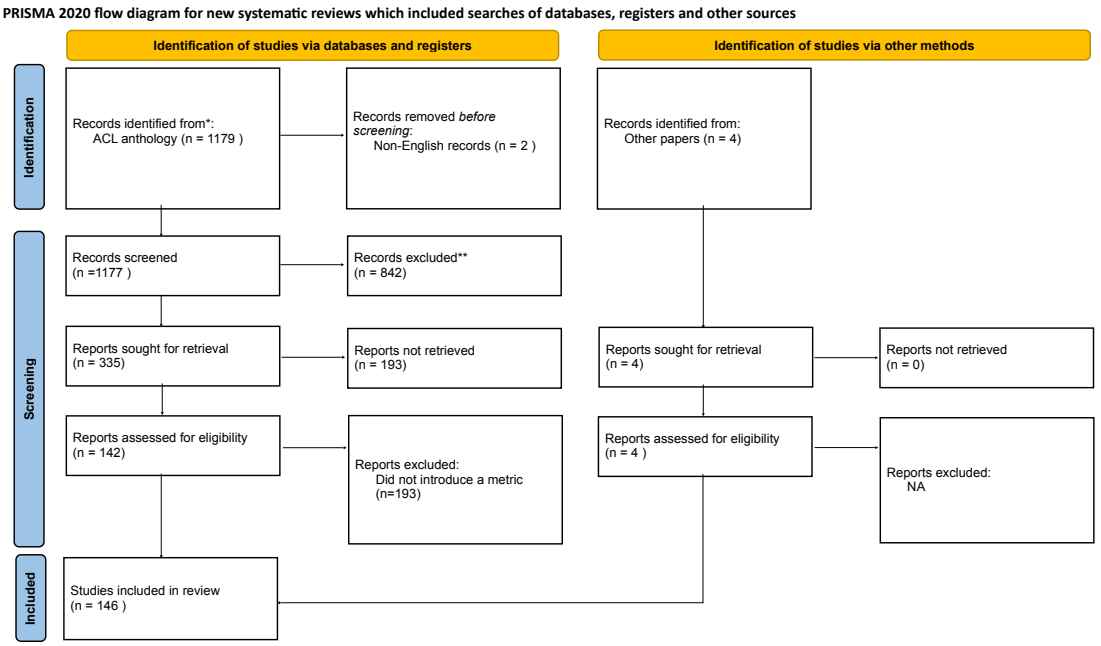

Figure 1: PRISMA 2020 flow diagram of our paper collection.

(2020); Mao et al. (2021); Basu Roy Chowdhury et al. (2021); Ciora et al. (2021); Dev et al. (2021); Bansal et al. (2022); Chalkidis et al. (2022); Jentzsch and Turan (2022); Krishna et al. (2022); Orgad et al. (2022); Savoldi et al. (2022); Smith et al. (2022); Wang et al. (2022); Adewumi et al. (2023); Deshpande et al. (2023); Jha et al. (2023); Ladhak et al. (2023); Lee et al. (2023); Luukkonen et al. (2023); Rao et al. (2023); Sabir and Padró (2023); Touileb et al. (2023); Triboulet and Bouillon (2023); Vargas et al. (2023); Yeh et al. (2023).