On Measures of Biases and Harms in NLP

Sunipa Dev[1*] **Emily Sheng**[2*] **Jieyu Zhao**[1*] **Aubrie Amstutz**[*]

Jiao Sun[3] **Yu Hou**[3] **Mattie Sanseverino**[1] **Jiin Kim**[1] **Akihiro Nishi**[1]

Nanyun Peng[1,3] **Kai-Wei Chang**[1]

[1]University of California, Los Angeles, [2]Microsoft Research, [3]University of Southern California

Abstract

Recent studies show that Natural Language Processing (NLP) technologies propagate societal biases about demographic groups associated with attributes such as gender, race, and nationality. To create interventions and mitigate these biases and associated harms, it is vital to be able to detect and measure such biases. While existing works propose bias evaluation and mitigation methods for various tasks, there remains a need to cohesively understand the biases and the specific harms they measure, and how different measures compare with each other. To address this gap, this work presents a practical framework of harms and a series of questions that practitioners can answer to guide the development of bias measures. As a validation of our framework and documentation questions, we also present several case studies of how existing bias measures in NLP—both intrinsic measures of bias in representations and extrinsic measures of bias of downstream applications—can be aligned with different harms and how our proposed documentation questions facilitates more holistic understanding of what bias measures are measuring.

1 Introduction

As language technologies and their applications become more widely deployed in our society, there are also increasing concerns of the disparate impacts and harms these technologies have on different demographic groups (Bolukbasi et al., 2016; Webster et al., 2018). To address some of these concerns, a large body of work has emerged to discuss (Gonen and Goldberg, 2019; Bender et al., 2021; Blodgett et al., 2021), detect (Bolukbasi et al., 2016; Nangia et al., 2020), measure (Caliskan et al., 2017; Zhao et al., 2019; Webster et al., 2018; Li et al., 2020), and mitigate (Dev and Phillips, 2019; Ravfogel et al., 2020; Sun et al., 2019; Dev et al., 2021a) the social biases encoded by NLP models.

Several of these works include bias measures comprising of metrics and datasets to define and investigate social biases within the constructs of a specific NLP task, such as text classification or machine translation. Though these works propose different approaches for measuring biases, there is often similarly a lack of explicit alignment to harms, as well as a lack of comparative understanding of the advantages and disadvantages between the different bias measures for various language tasks. As an example, for the task of coreference resolution, there are several measures investigating gender bias (Zhao et al., 2018; Rudinger et al., 2018; Lu et al., 2020; Webster et al., 2018; Cao and Daumé III, 2020). However, each measure is unique in either the targeted demographic groups, metrics, dataset sentence structures, or the definition of bias, all of which ultimately affect what harms are measured. A better understanding of bias measures ultimately enables better adaptation and deployment for specific use cases.

This paper is motivated by two main goals. The first goal is to define a practical framework for harms that is both theoretically-motivated and empirically useful for describing bias measures. We organize a framework that is motivated by concepts from social psychology and linguistics, and narrow down specific definitions and heuristics to tag normative notions of harm with which bias measures align. Moreover, we illustrate the utility of this measure-harm alignment exercise with case studies that demonstrate how a measure might unknowingly conflate different harms, or how separate measures with nearly identical task definitions can actually measure very different harms. The second goal is to define a collection of documentation questions around bias measures that helps others

[*]equal contribution

Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, pages 246–267
November 20–23, 2022.

capture measure limitations and align operationalizations of "biases" to harms. Documenting various attributes (e.g., considerations for targeted demographic groups and tasks, dataset limitations, bias metric definitions and motivations) of a bias measure can help practitioners better articulate harms, appropriate use cases, and limitations. To achieve these goals, we organize a practical framework of harms, a tagged collection of 43 existing bias measures and the associated harms, a set of documentation questions, and a collection of case studies.

2 Background

We clarify the definitions of several terms used throughout this paper.

Bias in NLP Bias in language models is commonly defined as "skew that produces a type of harm" (Crawford, 2017) towards different social groups, though it is a complex notion that is often not well-defined in existing literature (Blodgett et al., 2020; Delobelle et al., 2022; Talat et al., 2022). In the existing NLP literature, "biases" are often operationalized via a measurement model (Jacobs and Wallach, 2021) through bias measures. While these bias measures are proxies for evaluating bias, they are often necessarily localized to measuring very specific skews and lack context of how a system would be used by real users. Additionally, unstated assumptions and definitions often pervade these measures (Blodgett et al., 2021). It remains an open question whether these bias measures actually measure meaningful and useful distinctions of "biases"—this work provides initial explorations to answer this question for several measures.

Bias Measures Bias evaluations in NLP typically have been categorized broadly into *intrinsic* or *extrinsic* evaluations based on whether they measure biased associations within the word embedding spaces (Caliskan et al., 2017) or biased decisions from models for specific tasks (Mohammad, 2018; Webster et al., 2019), respectively. We define a *bias measure* as an evaluation standard that includes a *metric(s)* applied to a *dataset*. Here, we use the term *dataset* broadly, such that it could be applicable to datasets ranging in size and curation technique (e.g., manually crafted, generated). To show inequalities between demographic groups, existing works typically define bias metrics (e.g., specialized notions of group fairness) that they then apply to a dataset specially designed to reveal social inequalities or stereotypes.

These measures span several NLP tasks such as question answering (Li et al., 2020), relation extraction (Gaut et al., 2019), textual entailment (Dev et al., 2019), toxicity prediction (Dixon et al., 2018; Jigsaw, 2019; Sap et al., 2020), coreference resolution (Zhao et al., 2019; Cao and Daumé III, 2020), autocomplete generation (Sheng et al., 2019), dialogue generation (Dinan et al., 2019), machine translation (Stanovsky et al., 2019), as well as intrinsic measurements of the embeddings themselves (Caliskan et al., 2017; Bolukbasi et al., 2016; Lauscher et al., 2020; Malik et al., 2022).

Demographic Dimension We use the term *demographic dimension* to refer to an identity axis (e.g., gender, race, age) for which specific instances (e.g., for gender: *male*, *female*, *non-binary*, etc) are evaluated. Instances of a demographic dimension are typically comparatively evaluated in measures through some proxy, e.g., occupations or identity terms.

Harms While existing works have examined possible harms of NLP models from various perspectives (e.g., general social impacts (Hovy and Spruit, 2016), risks associated with large language models (Bender et al., 2021)), in the context of algorithmic biases, we seek to align specifically with harms that can arise specifically from biases. The relevant harms can be subdivided into representational or allocational harms, depending on whether there is a generalization of harmful representations of groups or if there is a tangible, disparate distribution of resources between groups, respectively (Crawford, 2017).[1] In the context of aligning bias measures with targeted representational harms, one could align with the motivations for creating the measure (either explicit or unstated), the techniques used, or some mix of both. Blodgett et al. (2020) present a categorization of the motivations and techniques of existing works that align with coarse-grained types of harms (*allocational*, *stereotypes*, *other representational harms*), and Blodgett (2021) further organize a taxonomy of fine-grained representational harm categories, including *quality of service*, *stereotyping*, *denigration and stigmatization*, *alienation*, and *public participation*. We build upon Blodgett (2021)'s discussions, framing and extending our curated framework of harms through documentation questions and heuristics that can

[1] Sheng et al. (2021) also separate out vulnerability harms, e.g., from model generations that render a group more susceptible to representational or allocational harms.

Task	Demographic Dimension	Bias Measure	Harms Evaluated
Coreference Resolution	**Gender** through identity terms	Webster et al. (2018) Cao and Daumé III (2020)	**QoS** **Erasure, QoS**
	Gender through occupations	Zhao et al. (2018) Rudinger et al. (2018) Lu et al. (2020)	**Erasure, Stereo.** **Erasure, Stereo.** **Erasure, Stereo.**
Natural Language Inference	**Gender** through occupations	Dev et al. (2019)	**Stereo.**
	Nationality through identity terms	Dev et al. (2019)	**Disparagement, Stereo.** through polar adj.
	Religion through identity terms	Dev et al. (2019)	**Disparagement, Stereo.** through polar adj.
Sentiment Analysis	**Age** through identity terms	Díaz et al. (2018)	**Disparagement, Erasure, QoS** through neg. sentiment
	Gender through identity terms	Kiritchenko and Mohammad (2018)	**Dehumanization, Erasure, QoS, Stereo.** through emotion words
	Rigid designators through references to specific people	Prabhakaran et al. (2019)	**QoS**
	Race through identity terms	Kiritchenko and Mohammad (2018)	**Dehumanization, Erasure, Stereo.** through emotion words
Question Answering	**Race** through identity terms	Li et al. (2020)	**Erasure, Stereo.** through neg. assoc.
	Ethnicity through identity terms	Li et al. (2020) Li et al. (2020) + Zhao et al. (2021)	**Erasure, Stereo.** through neg. assoc. **Erasure, Stereo.** through neg. assoc.
	Gender through occupations	Li et al. (2020) Li et al. (2020) + Zhao et al. (2021)	**Erasure, Stereo.** **Erasure, Stereo.**
	Religion through identity terms	Li et al. (2020) Li et al. (2020) + Zhao et al. (2021)	**Erasure, Stereo.** through neg. assoc. **Erasure, Stereo.** through neg. assoc.
Relation Extraction	**Gender** through hypernym (occupation) relation	Gaut et al. (2019)	**Erasure, Stereo.**
	Gender through spouse relation	Gaut et al. (2019)	**Erasure, Stereo.**
Text Classification	**Gender** through occupations	De-Arteaga et al. (2019) Zhao et al. (2020)	**Erasure, Stereo.** **Erasure, Stereo.**
	Gender through identity terms **Age** through identity terms **Region** through identity terms	Chalkidis et al. (2022) Chalkidis et al. (2022) Chalkidis et al. (2022)	**QoS** **QoS** **QoS**
Toxicity Detection	**Age** through identity terms	Dixon et al. (2018) Sap et al. (2020)	**Disparagement, Erasure** **Dehumanization, Disparagement, Erasure, Stereo.**
	Disability through identity terms	Dixon et al. (2018) Jigsaw (2019) Sap et al. (2020); Hutchinson et al. (2020)	**Disparagement, Erasure** **Disparagement, Erasure** **Dehumanization, Disparagement, Erasure, Stereo.**
	Gender through identity terms	Dixon et al. (2018) Park et al. (2018) Jigsaw (2019) Sap et al. (2020)	**Disparagement, Erasure** **Disparagement** **Disparagement, Erasure** **Dehumanization, Disparagement, Erasure, Stereo.**
	Rigid designators through references to specific people	Prabhakaran et al. (2019)	**QoS**
	Sexual Orient. through identity terms	Dixon et al. (2018) Jigsaw (2019) Sap et al. (2020)	**Disparagement, Erasure** **Disparagement, Erasure** **Dehumanization, Disparagement, Erasure, Stereo.**
	Race through identity terms	Dixon et al. (2018) Jigsaw (2019) Sap et al. (2020)	**Disparagement, Erasure** **Disparagement, Erasure** **Dehumanization, Disparagement, Erasure, Stereo.**
	Religion through identity terms	Dixon et al. (2018) Jigsaw (2019) Sap et al. (2020)	**Disparagement, Erasure** **Disparagement, Erasure** **Dehumanization, Disparagement, Erasure, Stereo.**
	Political Ideo. through identity terms	Sap et al. (2020)	**Dehumanization, Disparagement, Erasure, Stereo.**
	Victim through identity terms	Sap et al. (2020)	**Dehumanization, Disparagement, Erasure, Stereo.**

Table 1: Existing bias measures (part 1) organized by tasks, and demographic dimensions. A '+' indicates that one work built a bias metric (after '+') on top of a dataset from another work (before '+'). *Rigid designators*: references to specific people, *polar adjectives*: good vs bad; *negative activity*: violent or bad traits and activities. Sec. 5 delves into a few of these measures in the context of harms evaluated.

serve as a practical guide for those developing bias measures that capture specific harms.

Specifically, we use definitions of harms that are robust enough to capture aspects of a bias measure (dataset, metric(s), motivations) that align with different harms. Taking both individual and aggregate harms (Blodgett, 2021) into consideration, this framework assumes vulnerability to harm is mediated by a dominant—non-dominant identity group dichotomy (inspired by but not entirely aligned with Social Dominance Theory (Sidanius and Pratto, 2001)), which is helpful for operationalization purposes.

In this paper, we focus on five types of harm: Stereotyping, Disparagement,[2] Dehumanization, Erasure, and Quality of Service (QoS). While there are other types of harms, and the five we target could be further broken down into subcategories, we start with these five as they are previously studied concepts and provide interesting insights to the non-exhaustive list of bias measures we examine in Table 1 and Appendix Table 2.

3 A Framework for Harms

Conflating harms impedes accurate measurement; adequate and consistent delineation of harms enables ongoing appraisal of the effectiveness of mitigation strategies and the comparison of trade-offs. Our practical framework of harms builds upon existing taxonomies of representational harms (e.g. Blodgett (2021) and establishes specific heuristics (Appendix A) to disentangle the characteristics of five non-mutually exclusive categories. While these harms have previously been taxonomized, *we ground the definitions of harms into documentation questions and heuristics to help practitioners align NLP bias measures with specific harms*.

Addressing a single phenomenon with different lenses can surface multiple harms; precisely which harm a method captures is sometimes solely dependent on the experimental framing, rather than some inherent taxonomic difference. Using the harm heuristics we devise in Appendix A, we tag and distinguish between types of harms targeted by popular NLP bias measures presented in Tables 1 and Appendix Table 2. We note however, that other interpretations of targeted harms are certainly possible. This subjectivity makes it more crucial that those who build bias measures clearly state their motivations and include explanations of the relevant harms (Section 4).[3]

3.1 Harms

Stereotyping Stereotypes are overgeneralized beliefs about the personal attributes of an individual as determined by their demographic group membership. Stereotypes as entities are codified associations which are necessarily well-known within a given context (Devine, 1989) and can be expressed in infinite (and multi-modal) ways. Stereotypes draw on commonly held generalizations to make *a priori* judgements about groups. In human cognition, they are perpetuated through a process of discounting counter-evidence as exceptions to the rule, e.g. confirmation bias (Allport et al., 1954; Link and Phelan, 2001). These associations can in turn lead to unintended "affective reactions" by the model—precisely the measurable signals from which practitioners can infer bias.

Disparagement Disparagement encapsulates any behavior by a model which reinforces the notion that certain groups are less valuable than others and less deserving of respect (or resources). Commonly associated measures of disparagement include toxicity ratings and hate speech detection scores.

Dehumanization Dehumanization actively casts disfavored groups as "others" and aims to erase signs of shared humanity (e.g. emotions, agency, intelligence), thus suppressing opportunities for empathy with said "out group" by characterizing them as sub-human (Markowitz and Slovic, 2020; Haslam and Stratemeyer, 2016). Dehumanization can therefore be challenging to measure directly, as instances of dehumanizing language or sentiments are often closely intertwined with Disparagement and Stereotyping.

Erasure Erasure refers to the lack of adequate representation of members of a particular social group (Dev et al., 2021b; Blodgett et al., 2022), whether intentional or not. While the data used to represent the intricacies of reality will always be necessarily incomplete, Erasure can arise from mismatches in reality and the data chosen to represent it. It can also serve to reinforce existing power structures via incautious mathematical averaging or aggregation of disparate groups. While

[2] We choose to use "Disparagement" instead of "Denigration", to avoid invoking the conceptual metaphor of 'blackening' one's reputation, which can have racial connotations in US culture.

[3] We also note that it is sometimes difficult to align with certain harms like Dehumanization without a closer examination of all samples in a measure dataset.

relational group sizes from the real world can be reflected from the model in a quantitative sense, the challenge is designing systems which do not allow relative size to inappropriately affect prominence, i.e., attention needs to be paid to the potential effects these probabilities have on produced output.

Quality of Service Quality of Service harms result from instances where a model fails to perform equitably for different groups (Blodgett, 2021). This harm can in turn potentially result in inequitable allocation of resources (Blodgett et al., 2020), though this harm can also exist independently. The potential 'quality' of service is operationalized and quantified via defined performance indicators, which can be systematically compared between commensurable groups.

3.2 Relationships between Harms

Disentangling which categories of harm a given bias measure measures requires careful articulation of the hypothesis and documentation of operationalization decisions; framing is crucial for producing substantively valid results (Jacobs and Wallach, 2021). For example, an instance of bias in model training data may have arisen due to multiple types of harm or cause multiple types of harm. Our framework emphasizes how consequential these distinctions in operationalization can be.

Disparagement and Stereotyping Because stereotypes need to be codified and well-known within a given culture (Devine, 1989), Disparagement is more generic and group-agnostic than Stereotyping. Consequently, datasets that test for Disparagement (explicitly or not) may sometimes be generated *ad infinitum* by swapping demographic identifiers, e.g., "[demographic identifier] are the worst kind of people". In comparison, the specificity required of statements expressing stereotypes presents limitations on rephrasing concepts (by design, languages have few "absolute synonyms" (Murphy, 2010)).

Dehumanization, Disparagement, and Stereotyping Under our framework, Dehumanization contributes to Disparagement because it reinforces the idea that certain groups are inherently less valuable to society, i.e., Dehumanization always serves Disparagement, but not *vice versa*. Dehumanizing language uses techniques such as *moral disgust, denial of agency,* or *likening members of a target group to non-human entities* (Markowitz and Slovic, 2020) to reinforce normative identities—often as indication of a biological hierarchy of 'species' within humankind. Dehumanization can be "expressed tacitly" (Markowitz and Slovic, 2020), e.g., when groups are not considered worthy of being included (via Erasure). While descriptive, proscriptive, or prescriptive stereotypes (Koenig, 2018; Hall et al., 2019) may have originated from some quantitative or qualitative fact about societal norms (Sidanius and Pratto, 2001), stereotypes which dehumanize are more likely inherently unfounded, e.g., stereotypes perpetuating racist pseudoscience like eugenics.

Stereotyping and Erasure Cognitive heuristics like categorization and prediction based on probability are part of human nature (Tversky and Kahneman, 1974; Mervis et al., 1981); however, harm can arise when these associations obfuscate or erase actual variance (e.g., via confirmation bias) or when society assigns a cost (e.g., social, allocational) when these oversimplified "norms" are not adhered to by their respective group members (e.g., proscriptive or prescriptive (Koenig, 2018)). Erasure and Stereotyping can have a cyclical relation; lack of representation of variance and sub-populations can both result in stereotypes and be a direct result of Stereotyping. Erasure and Stereotyping are conceptualized as being one level of abstraction away from the consequence being caused: while exposure to a disparaging or dehumanizing remark can be directly harmful in the moment, the impact of Stereotyping associations and Erasure are more apparent at a distributional level. Additionally, Erasure and Stereotyping are strongly mediated by the *vulnerability* of the group and the *severity* of the implications of the association.

Quality of Service and Erasure Facts about historical inequities, social hierarchies, and stereotypes should guide Erasure measures. Under our framework, measures that target Erasure harms should have strong, directional hypotheses in order to surface representation issues for specific groups. These issues could in turn be quantified more precisely via comparative evaluation methods, such as those common in measures that target Quality of Service harms. Erasure measurement for underrepresented groups requires us to set aside quantitative majorities and ensure qualitative "coverage" instead, e.g., while there may be fewer female than male surgeons in the United States, the former do still exist. The desired effect of removing Erasure harms is for representation of actual diversity to

persist, independent of statistical presence.

4 Documenting Bias Measures

While bias measures aimed at various tasks are widely developed across the NLP community, the measures are often underused or re-developed by researchers for the same task. This stems largely from a lack of usability since little to no documentation of motivation and various choices is available for these measures. Documentation for datasets and models have proliferated over the last few years but the rapidly growing collection of bias measures lacks such organized efforts.

Existing works have stressed the importance of documenting models (Mitchell et al., 2019), datasets (Gebru et al., 2018; Bender and Friedman, 2018), measurement modeling validity and reliability (Jacobs and Wallach, 2021), and, more recently, ethical considerations (Mohammad, 2022). This paper adds a complimentary resource focusing on documentation considerations for bias measures into the existing collection. In this work, we build upon the existing guidelines from Gebru et al. (2018), which are more generally for datasets of any modality or purpose, and narrow the focus to bias measures for NLP tasks. We add questions related to the *Composition* and *Collection Process* sections as proposed by Gebru et al. (2018). Additionally, we propose new sections on *Motivation* specifically for bias measures and *Bias Metrics*. The specificity of the questions addresses the intended usage of different bias measures more explicitly.

1. *Motivation*
   Blodgett et al. (2020) detail the importance of concretely defining the biases being measured and listing out how a metric aligns with normative definitions of harm. Additionally, discerning biases from model errors is equally important and particularly ambiguous when a definition for the "bias" measured is absent.
   - **What is the stated definition of bias?**
   - **How does this definition align with normative definitions of harm?** For a measure to be a valid quantification of bias, the notion of "bias" has to be well-defined and related to what is measured. More explicitly bridging the gap between bias metrics and harms can tangibly disambiguate between innocuous model errors and potential harms downstream.
   - **If the bias measure measures more than one harm, are the harms conflated in one measurement or separable?** A single instance of language may represent/cause multiple forms of harm (e.g., some Stereotyping harms may also be Dehumanization harms). Does the measure provide a method for measuring multiple harms separately as well as in aggregate (e.g., are subsets of the underlying data tagged along multiple axes)?
   - **What language and culture is the bias and measure most relevant to?**
   - **What other contexts can the measure be extended to?** This question is intended to obtain a list of the specific demographic groups and locales a bias measure has been shown to be useful for.
   - **If a demographic attribute is split into groups for measurement of bias, how many groups have been considered?** What is the justification for the grouping? Have prominent/consequential intersectional identities been considered? This question is to understand the scope of the measure and assess its coverage.
   - **What is the source of bias that is measured?** Social biases creep into NLP models in different ways - the data used to derive representations, the model (and parameters) used, etc. The bias measured can be from one or all sources and needs to be acknowledged and when possible, disambiguated.
   - **What tasks or applications is this bias measure useful for?** Is this measure effective to check on any language representations for social biases irrespective of application? Or is there a specific task where this is most applicable?
2. *Composition and Collection Process*
   Language data for bias measures is sourced primarily in two ways: by extracting from existing textual data or by generating from specific templates. While the first has the advantage of being more similar to "real samples" that models see, the latter has the advantage of testing for specific artifacts by construct.
   - **Is the bias measure data scraped, generated, or produced some other way?** Scraping or generating text using templates are two common ways of building bias measure datasets in NLP, and different dataset curation techniques have their own advantages and disadvantages.

- **What are the limitations associated with method of data curation? How generalizable is this dataset?** Examples of limitations include scraped English text containing predominantly Western narratives and data annotated by annotators with specific biases.
- **If the dataset is scraped, what are the primary sources/domains?** Some text sources are known to harbor more toxic or harmful content than others.
- **What is the structure of the sentence, sentence segment, template, or trigger phrase used for data collection?** Does the particular structure come with certain simplifications, assumptions, or guarantees?
- **Is the dataset at risk of causing harm through the particular selection of proxy attributes representing demographic groups?** For example, does this dataset use popular names as a proxy for gender? Is there a risk for misidentifying individuals if the associated genders are not self-reported? Does the expected gender - name pairing align with the time period of the sourced data?

3. *Bias Metrics*

   This section presents documentation questions for metrics that are used with datasets to measure bias. Specific definitions and comparisons can broaden understanding about the measured biases.

   - **How is the bias metric defined? Is there a null hypothesis or normalization recommended for it to be meaningful?**
   - **Is it an absolute or relative evaluation?** Sheng et al. (2021) describe absolute score evaluations as those that "use an accumulated score to summarize inequalities between demographics, whereas relative evaluations explicitly report inequalities between all demographics." Absolute scores offer more simplicity, and relative scores offer more flexibility in alignment with normative harms. Through this question, we hope to understand the motivation behind the evaluation format.
   - **Are there alternate or existing metrics this metric can or should be used with?** This question covers the cases where a bias metric may not be enough to measure all desired metric attributes, either in terms of bias or general task evaluations.
   - **Are there other existing datasets or metrics to evaluate bias for the same task?** How does an evaluation using one metric correlate with another using a different metric? Note that high correlation between measures do not necessarily imply meaningful or useful measures. Additionally, does the sentence structure, sourcing method, or other feature differ between the datasets?
   - **Can the metric imply an absolute absence of bias in a specific task or model?** Are there other measurements needed for a complete assessment of bias? Is a complete assessment possible?

5 Case Studies

We present a series of case studies as examples of how our proposed framework of harms and documentation questions reveal unique insights into different bias measures. In Table 1 and Appendix Table 2, we tag bias measures with the relevant, targeted harm(s). In this section, we discuss concrete examples to elucidate how subtle differences in framing of measures impact the harm(s) measured.

5.1 Disparagement and Stereotyping

To better understand the subtleties between Disparagement and Stereotyping, we examine two existing bias measures.

Davani et al. (2020) present a fair hate speech measure that implicitly separates Stereotyping and Disparagement harms; however, these alignments are not explicitly connected, and our framework helps distinguish between the two harms. This work of Davani et al. (2020) is motivated by the observation that not all demographic groups are interchangeable when it comes to specific stereotypes. For example, they note that substituting "Muslim" with "Jew" in a hateful sentence about terrorism does not create equivalently valid stereotypes within the US cultural context. Thus, they create "symmetric counterfactual" statements that convey a similar meaning when different group tokens are substituted. Interestingly, this distinction between symmetric and asymmetric counterfactuals helps delineate between Disparagement and Stereotyping sentences, as symmetric counterfactuals are, by nature, generic enough to disparage multiple groups. Unless two independent stereotypes have coincidentally converged (e.g., two groups are associated with terrorism for different historical reasons within a given context), a carrier phrase

that is able to substitute group identifiers is unlikely to be able to produce valid stereotypical sentiments. Thus, this process of creating and making the distinction between symmetric and asymmetric counterfactual tests generates a fair hate speech dataset that includes some amount of coverage for both Disparagement and Stereotyping harms.

Dev et al. (2019) is another example where Disparagement and Stereotyping harms are not explicitly separated. This work measures biases in the task of natural language inference by comparing demographic associations with polar adjectives. We find that this particular setup conflates Disparagement and Stereotyping harms. As an example from the dataset, for the template "[demographic identifier] are [adjective]", the statement "Canadians are nice" is a stereotype, whereas another statement such as "Uzbekistanis are bad" is more of a general disparaging remark than a stereotype.

These examples show the difficulty in carefully designing datasets that test for Stereotyping versus Disparagement harms.

5.2 Quality of Service, Stereotyping, and Erasure

Next, we present an empirical case study examining how measures designed for the same task can differ in the harms measured. Webster et al. (2018) and Cao and Daumé III (2020) both discuss biases in the task of coreference resolution where the goal is to identify phrases or terms referring to the same entity in a sentence. Webster et al. (2018) measure biases in the model's ability to correctly resolve gendered pronoun-name relationships for the binary genders and is aligned with the Quality of Service harm, since the measure probes the contrastive relationship between model performance for females versus males. Cao and Daumé III (2020) expand the `GAP` dataset introduced by Webster et al. (2018) to create the `MAP` dataset, where the authors swap out gendered words for a set of gender neutral variations of the sentences in `GAP`. While both `GAP` and `MAP` are part of bias measures that are aligned with Quality of Service harms, `MAP` also surfaces Erasure harms by testing for whether a coreference system fails to process text for non-binary pronouns.

Additionally, two other popularly used bias measures for coreference resolution, as described by Rudinger et al. (2018) and Zhao et al. (2019), compare the association of specific occupations with gendered pronouns. While some dataset instances directly measure Stereotyping harms, such as a preferential association of 'doctor' with typical male pronouns, other instances do not directly measure explicit stereotypes in the society but rather an implicit Erasure or lack of representation of some genders in overall text. While both of these harms are overall conflated by both measures, unlike `GAP` and `MAP`, neither measures Quality of Service harms.

5.3 Dehumanization and Stereotyping

Kiritchenko and Mohammad (2018)'s bias measure for sentiment analysis formulates a dataset of simple sentences including names, gendered pronouns, and other indicators of demographic group identity, and compares the sentiment associated with different groups. While some sentences evaluate stereotypes such as the "Angry Black Woman", others are not indicative of any stereotype but rather analyze the societal license for a member of a certain group to display a range of emotions–i.e., Dehumanization. The two harms measured are not distinguishable by the metric used, but instead by careful examination of the individual sentence templates, word lists, and names used.

5.4 Insights from Documenting Bias Measures

By using our harm framework to label the bias measures in Table 1 and Appendix Table 2 as well as documenting bias measure motivations and compositions, we developed several insights.

The first is that *documentation facilitates deeper analysis and should be revisited periodically*. We use the proposed questions to analyze the work described by Sheng et al. (2019). In particular, we note that there is no explicit definition of biases in the work, although the operationalization of their regard metric as a measure of social perception aligns with the measurement of representational harms (e.g., Stereotyping and Disparagement). In answering the documentation questions (Appendix C.2), we find that this documentation exercise is especially useful if the documented measure has been released for a while. In the case of the regard metric of Sheng et al. (2019), there were not many points of comparison at the time of its release, but more relevant comparisons have recently been released. Thus, we recommend treating documentation as a continuous process and revisiting the questions regularly.

Also, *documentation reveals specific limitations across bias measures for a specific task*. The specificity of the documentation questions helps uncover what is currently measured and encourages the development and use of complementary measures. In documenting WinoBias (Zhao et al., 2018) in Appendix C.1, we examine various bias measures for coreference resolution more closely. Existing bias measures for coreference resolution that target gender biases through occupations have all focused on associated stereotypes and the relative representation between binary genders, and thus target Stereotyping and Erasure harms, as shown in Table 1. On the other hand, the coreference resolution bias measures that target gender through identity terms explore the effect of model performance for gender-neutral pronouns, and thus target Quality of Service (and some Erasure) harms.

A third insight is that *inherent constraints of a task seem to affect the method by which bias measures (implicitly or explicitly) target harms*. For more constrained language understanding tasks in which the model produces a limited set of outputs (e.g., classification), the dataset designed for the measure largely affects the targeted harm. For example, for measuring biases in coreference resolution, the standard metrics are F_1 or accuracy scores—it is really by examining the datasets (and motivations) that we discern whether we are targeting Stereotyping (e.g., through occupational associations) or Quality of Service harms. For open-domain language generation tasks, targeted harms are largely affected by the selected bias metrics rather than the datasets. Because generation task are so open-ended, it is often difficult to design evaluation datasets that achieve a lot of control over the resulting model output, and thus existing works rely more on various bias metrics to capture different harms. For example, Dhamala et al. (2021) evaluate biases using sentiment, regard, toxicity, and psycholinguistic norms to target different operationalizations of harms.

6 Conclusion

Bias measures in NLP are critical for estimating and mitigating potential harms towards different demographic groups. However, a lack of structured understanding of what harms exist, how they are operationalized through bias measures, and how they can be measured can diminish the usefulness of bias measures. In this work, we organize a framework to define and distinguish between different types of harms—presented through heuristics and documentation questions—to guide more intentional development of bias measures. Our proposed documentation template also facilitates combining, comparing, and utilizing different bias measures, and continuously re-visiting them to update limitations and comparative understanding with other measures.

7 Limitations and Ethical Considerations

We acknowledge that our framework of harms has been created from a US-centric perspective and has been influenced by the Social Dominance Theory (Sidanius and Pratto, 2001), which can be limiting from a global perspective and does not include cultural harms. While some definitions and operationalizations of harms in our framework (e.g., Stereotyping, Disparagement) may be applicable to other cultural perspectives, we note that there may be some that require cultural context-specific updates and also that there are other harms that we did not cover. There are also other bias measures in this rapidly growing space that we may not have covered and tagged with harms measured. Additionally, we do not focus on specific downstream applications where each measure might be used and encourage further analysis on these applications.

We further emphasize that while documentation enables more transparency into bias measures, documentation *does not ensure the validity* of the measures. In fact, there is a risk that the act of documentation could give a measure a false sense of validity. Too many documentation questions may also become an obstacle for practitioners interested in working on a topic, though we believe it is better for community progress to start thinking about these questions before designing bias measures.

Acknowledgments

We would like to thank various people for their valuable discussion and feedback, including Alexandra Olteanu, Arjun Subramonian, Chad Atalla, Dan Vann, Emily Corvi, Hanna Wallach, Hannah Washington, Jason Teoh, Kevin Robinson, Stefanie Reed, Su Lin Blodgett, Vinodkumar Prabhakaran, as well as our anonymous reviewers. This work was supported by NSF grant #1927554, NSF grants #2030859 and #2127309 to the Computing Research Association for the CIFellows Project, and the Sloan Award.

References

Gordon Willard Allport, Kenneth Clark, and Thomas Pettigrew. 1954. The nature of prejudice.

Sarah Alnegheimish, Alicia Guo, and Yi Sun. 2022. Using natural sentences for understanding biases in language models. *NAACL*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big. *Proceedings of FAccT*.

Su Lin Blodgett. 2021. Sociolinguistically driven approaches for just natural language processing.

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*.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of african-american english. *arXiv preprint arXiv:1608.08868*.

Su Lin Blodgett, Q. Vera Liao, Alexandra Olteanu, Rada Mihalcea, Michael Muller, Morgan Klaus Scheuerman, Chenhao Tan, and Qian Yang. 2022. Responsible language technologies: Foreseeing and mitigating harms. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, New York, NY, USA. Association for Computing Machinery.

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 *Proc. 59th Annual Meeting of the Association for Computational Linguistics*.

T Bolukbasi, K W Chang, J Zou, V Saligrama, and A Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *ACM Transactions of Information Systems*.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500.

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

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Felix Schwemer, and Anders Søgaard. 2022. Fairlex: A multilingual benchmark for evaluating fairness in legal text processing. *arXiv preprint arXiv:2203.07228*.

Kate Crawford. 2017. The trouble with bias. Keynote at NeurIPS.

Aida Mostafazadeh Davani, Ali Omrani, Brendan Kennedy, Mohammad Atari, Xiang Ren, and Morteza Dehghani. 2020. Fair hate speech detection through evaluation of social group counterfactuals. *arXiv preprint arXiv:2010.12779*.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

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.

Sunipa Dev, Tao Li, Jeff Phillips, and Vivek Srikumar. 2019. On measuring and mitigating biased inferences of word embeddings.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. 2021a. 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, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021b. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *EMNLP*, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sunipa Dev and Jeff Phillips. 2019. Attenuating bias in word vectors. In *AISTATS*, Proceedings of Machine Learning Research, pages 879–887. PMLR.

Patricia G Devine. 1989. Stereotypes and prejudice: Their automatic and controlled components. *Journal of personality and social psychology*, 56(1):5.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *FAccT*.

Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. 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.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. 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 4537–4546.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Aparna Garimella, Carmen Banea, E. 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 *ACL*.

Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth M. Belding, Kai-Wei Chang, and William Yang Wang. 2019. Towards understanding gender bias in relation extraction. *CoRR*, abs/1911.03642.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, and Derek Michael Wu. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, page 229–233, New York, NY, USA. Association for Computing Machinery.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sanchez, Mugdha Pandya, and Adam Lopez. 2020. Intrinsic bias metrics do not correlate with application bias. *arXiv preprint arXiv:2012.15859*.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of NAACL-HLT 2019*, pages 609–614. Association for Computational Linguistics.

Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. 2020. Investigating African-American Vernacular English in transformer-based text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5877–5883.

Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic gender identification and reinflection in Arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165.

Erika V Hall, Alison V Hall, Adam D Galinsky, and Katherine W Phillips. 2019. Mosaic: A model of stereotyping through associated and intersectional categories. *Academy of Management Review*, 44(3):643–672.

Nick Haslam and Michelle Stratemeyer. 2016. Recent research on dehumanization. *Current Opinion in Psychology*, 11:25–29. Intergroup relations.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

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.

Jigsaw. 2019. *Jigsaw Unintended Bias in Toxicity Classification*.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442.

Svetlana Kiritchenko and Saif M. Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems.

Anne M Koenig. 2018. Comparing prescriptive and descriptive gender stereotypes about children, adults, and the elderly. *Frontiers in psychology*, 9:1086.

Anne Lauscher, Rafik Takieddin, Simone Paolo Ponzetto, and Goran Glavaš. 2020. AraWEAT: Multidimensional analysis of biases in Arabic word embeddings. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 192–199.

Tao Li, Tushar Khot, Daniel Khashabi, Ashish Sabharwal, and Vivek Srikumar. 2020. Unqovering stereotyping biases via underspecified questions.

Bruce G. Link and Jo C. Phelan. 2001. Conceptualizing stigma. *Annual Review of Sociology*, 27(1):363–385.

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

Haochen Liu, Wentao Wang, Yiqi Wang, Hui Liu, Zitao Liu, and Jiliang Tang. 2020b. Mitigating gender bias for neural dialogue generation with adversarial learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 893–903.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*. Springer.

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.

David M. Markowitz and Paul Slovic. 2020. Social, psychological, and demographic characteristics of dehumanization toward immigrants. *Proceedings of the National Academy of Sciences*, 117(17):9260–9269.

Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, and A. Galstyan. 2020. Man is to person as woman is to location: Measuring gender bias in named entity recognition. *Proceedings of the 31st ACM Conference on Hypertext and Social Media*.

Ninareh Mehrabi, Pei Zhou, Fred Morstatter, J. Pujara, Xiang Ren, and A. Galstyan. 2021. Lawyers are dishonest? quantifying representational harms in commonsense knowledge resources. *ArXiv*, abs/2103.11320.

Carolyn B Mervis, Eleanor Rosch, et al. 1981. Categorization of natural objects. *Annual review of psychology*, 32(1):89–115.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184.

Saif Mohammad. 2022. Ethics sheets for AI tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8368–8379, Dublin, Ireland. Association for Computational Linguistics.

R. Munro and Alex Morrison. 2020. Detecting independent pronoun bias with partially-synthetic data generation. In *EMNLP*.

M. Lynne Murphy. 2010. *Lexical and semantic relations*, Cambridge Textbooks in Linguistics, page 108–132. Cambridge University Press.

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)*, 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.

Aurelie Neveol, 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.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804.

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.

Daniel Preoţiuc-Pietro and Lyle Ungar. 2018. User-level race and ethnicity predictors from Twitter text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1534–1545, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in Elicited Natural Language Inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *NAACL*, pages 8–14.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678.

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.

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.

Emily Sheng, Kai-Wei Chang, Premkumar 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*.

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. 2020. "you are grounded!": Latent name artifacts in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6850–6861.

J. Sidanius and F. Pratto. 2001. *Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression*. Cambridge University Press.

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.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode #5 – Workshop*

on Challenges & Perspectives in Creating Large Language Models, pages 26–41, virtual+Dublin. Association for Computational Linguistics.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

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.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Kellie Webster, Marta R. Costa-jussà, Christian Hardmeier, and Will Radford. 2019. Gendered ambiguous pronoun (GAP) shared task at the gender bias in NLP workshop 2019. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 1–7.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14.

Jieyu Zhao, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Kai-Wei Chang. 2021. Ethical-advice taker: Do language models understand natural language interventions? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4158–4164.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender Bias in Contextualized Word Embeddings. In *Proceedings of NAACL-HLT 2019*, pages 629–634. 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.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661.

Appendix: On Measures of Biases and Harms in NLP

A Harm Framework Heuristics

To help practitioners determine the specific harm(s) a bias measure evaluates, we propose the following set of heuristics.

Stereotyping: Does the method:

- deal with language which communicates an existing, well-known a priori judgement or generalization which oversimplifies the reality of diversity within the group?
- measure predictions or probabilities of associations between specific groups and certain characteristics, concepts, language, or sentiments?
- focus on finding specific, pre-defined outcomes based on hypotheses about stereotypical associations, i.e., is the hypothesis directional?
- test associations which either the "average" in-group member or person in the relevant society would be able to quickly predict, i.e., would they be able to predict or identify what the 'problem' is and connect its roots to their cultural/historical knowledge?

Note: these associations can be positive or negative, but should not hold as naturalistic when a commensurable group is swapped in.

Erasure: Does the method:

- search for lack of representation of specific groups based on cultural trends and patterns of historical inequality?
- engage with mismatches between representation and reality (due to imprecise categorizations, rounding errors, etc.)?
- interrogate representation issues caused by prevailing stereotypes, dehumanization, or cultural narratives?
- primarily concern itself with whether or how specific, pre-defined groups are represented or treated equitably, rather than to what extent groups are treated inequitably in relation to one another?
- primarily provide results about the model performance for a specific group in relation to a 'control' group (whether or not explicitly stated as such)?

Disparagement: Does the method:

- deal with generally belittling, devaluing, or de-legitimizing language about a group?
- engage with sentiments related to societal regard (respect), expressing normative judgments, or using scalar adjectives pertaining to quality or worth (best/worst, good/bad), but which are not tied to an established stereotype?
- use language which holds as pragmatically and semantically valid/naturalistic when the group identifier is perturbed with a commensurable group?
- deal with 'toxicity' or 'unhealthy' discourse in general?

Dehumanization: Does the method specifically mention language commonly used to dehumanize, such as:

- associations with non-human life (vermin, insects)?
- implications that a certain group is sub-human or not 'true' members of a superset (certain 'immigrants' aren't 'American')?
- notions related to eugenics?
- justifications of inequitable treatment of groups or denial of human rights based on group membership (note: these can be codified into stereotypes, but are distinguished by their unique purpose to 'other' the group, reinforcing notions of normative identities and casting divergence as indication of a hierarchy of 'species' within humankind)?

Quality of Service: Does the method:

- seek to measure the comparative performance of a model for several commensurable demographic groups?
- have an obvious or direct application to mitigation efforts or industry usage?
- primarily concern itself with to what extent groups are treated inequitably (quantification), rather than whether they are treated differ-

ently?

B A Survey of Bias Measures for Understanding Harms

As NLP models grow in size, complexity, ability to mimic underlying languages, and the extent to which they are deployed in real world applications, it becomes more important to understand their potential for biases and harms. A growing number of measures serve to evaluate biases in tasks such as sentiment analysis or relation extraction, targeting specific social biases related to gender, race, religion, etc. While measures to evaluate biases have been formulated across various tasks, there remains a lack of cohesive understanding of *what these bias measures evaluate* and *how different measures relate*. In this section, we survey and describe a non-exhaustive list of measures for quantifying biases in different NLP tasks for primarily English. Tables 1 and 2 summarize this survey along with alignments of harms for different bias measures.

B.1 Natural Language Understanding

We discuss existing works that use different measures to assess the presence of social biases in a variety of NLU tasks.

Coreference Resolution Coreference resolution is the task of finding all expressions that refer to the same entity in text; a more specific objective is to associate pronoun mentions to different entities. There are two distinct definitions of bias that are evaluated with respect to this task, both centered around gender. The first defines bias as model performance discrepancy across different groups of a demographic attribute like gender. The Gendered Ambiguous Pronouns (GAP) dataset (Webster et al., 2018) consists of samples from Wikipedia biographies with ambiguous pronoun-name resolution pairs. Webster et al. (2018) defines and measures biases through a disparity in correctly resolving pronoun-name relationships for the male and female genders. The Maybe Ambiguous Pronoun (MAP) dataset (Cao and Daumé III, 2020) expands GAP to go beyond binary genders with a broader dataset. The second category of coreference resolution bias measures investigates the propagation of stereotypes from language representations used by models. Both WinoBias (Zhao et al., 2018) and Winogender (Rudinger et al., 2018) generate Winograd schema style datasets to investigate occupational gender stereotypes. Additionally, Lu et al. (2020) create sentence templates to evaluate biases using the ratio of accurate pronoun resolution for stereotypical vs non-stereotypical occupational associations.

Existing works that use the second definition of bias currently focus on singular stereotypes (e.g., with regards to occupation), while gender biases can encompass a broad range of other stereotypical and undesired associations. While both definitions of bias can potentially cover additional demographics and undesired associations, it is important to question which is more applicable to investigate harms faced by a group. For example, non-binary individuals face erasure in language representations (Dev et al., 2021b), and these experienced harms might be more appropriately captured by the first definition, whereas stereotyping might be by the second.

Natural Language Inference (NLI) NLI determines the directional relationship between two sentences, as to whether the second sentence (hypothesis) is entailed, contradicted, or neutral to the first sentence (premise). Dev et al. (2019) demonstrate how the task captures and mirrors stereotypical associations (with binary gender, religion, etc) learned by text representations. Their bias measure consists of a dataset with sentence pairs: one sentence with an explicit demographic attribute (e.g., gender), and the other with implicit, stereotypical associations (e.g., occupations). Bias is measured as the accuracy of models in identifying that all sentences have no directional relation, i.e., classified as having the 'neutral' label. Since an overall bias score is calculated over a set of templates, a variety of templates can be independently assessed together to evaluate bias of NLI model outcomes across multiple demographic groups, thus not restricting measurements to a single stereotype.

Sentiment Analysis Estimating the sentiment or language polarity of text is useful for understanding consumer perception from reviews, tweets, etc. However, this task has been demonstrated to be stereotypically influenced by demographic characteristics such as race and gender (Kiritchenko and Mohammad, 2018), age (Díaz et al., 2018) and names of individuals (Prabhakaran et al., 2019). Existing works keep sentence templates constant between samples and change the assumed demographic attribute of the person (e.g., through names) in a sentence. This ideally should not change the sentiment classification of the sample—any

changes in sentiment indicate the existence of stereotypical associations. Since evaluation hinges on this contrast in classification across groups, bias against a group is also measured in comparison to another.

Question Answering (QA) QA models perform reading comprehension tasks and also propagate stereotypical associations from underlying language representations, as demonstrated through UnQuover (Li et al., 2020). Li et al. (2020) use sentence templates containing limited direct demographic information (e.g., names) and underspecified questions containing no related demographic information to measure biases exhibited by QA systems. The setup is such that all subcategories of a demographic attribute (e.g., religion: Christian, Buddhist, etc) should be equally predicted as the answer. A statistically significant, higher value for one sub-category is interpreted as bias. Thus, this measure expands the understanding of comparative biases across several demographic dimension values and is a closer reflection of the complexities of real-world biases.

Neural Relation Extraction Relation extraction is the task of extracting relations between entities in a sentence and is instrumental in converting raw, unstructured text to structured data. Gaut et al. (2019) note how gender biases in this task could lead to allocational harms by affecting predictions on downstream tasks. They create a dataset, WikiGenderBias, containing sentences regarding either a male or female entity and one of four relationships: spouse, occupation, birth date, or birth place. Similar to GAP, the evaluation framework measures gender bias as a difference in model performance for each gender. Instead of overall performance, they average over individual groups within a relationship (e.g., different individual occupations). This measure faces the challenge of generalizability as it relies on scraping a variety of existing text for different demographic groups.

Masked Language Model Predictions Several language representations are trained on the ability to predict masked words in text. CrowS-Pairs (Nangia et al., 2020) and StereoSet (Nadeem et al., 2021) are datasets that use this property to expose and evaluate social biases learned with respect to different protected attributes. Both use crowdsourcing to obtain annotated sentence pairs, one of which is more stereotypical than the other for specific attributes (gender, socioeconomic status, etc). The evaluation metrics in both measures grade the model on its preference (through probabilities) for either the stereotypical or other sentence. Because these datasets permit crowdworkers to provide free-flowing text, the datasets are able to expand understandings of biases beyond a single stereotypical association across groups.

Text Classification (Occupations) De-Arteaga et al. (2019) set up a measure for evaluating bias in text classification where the task is to predict a person's occupation given their biography.The dataset contains short biographies crawled from online corpora using templates and removing sentences which contain occupation names. Bias is evaluated by comparing results across different gender groups. Zhao et al. (2020) extend the original dataset to Spanish, French, and German. A challenge is equally scraping diverse data for different demographics, as reflected in the focus on binary gender for this measure.

Toxicity Detection Toxic language ranges from more explicitly offensive forms (e.g., vulgar insults) to more subtle forms (e.g., microaggressions). While toxicity detection aims to identify toxic language, existing works have found uneven detection of toxic language towards different groups. Prabhakaran et al. (2019) show that there are varying levels of toxicity towards different names. Dixon et al. (2018) analyze biases in a toxicity classification model through the Wikipedia Talk Pages dataset as well as through a templated test set. Jigsaw (Jigsaw, 2019) contains comments from the Civil Comments platform labeled with six types of toxicity (e.g., toxic, obscene, etc) and identity attributes (e.g., white, woman, etc). Along with this dataset, Jigsaw (2019) present a bias evaluation following that of Borkan et al. (2019) by comparing the AUC scores from different subgroups. Additionally, Sap et al. (2020) create a social bias inference corpus with toxicity labels and targeted group labels to understand the bias implications in languages. These bias measures demonstrate that even tasks intended to detect harms may be biased.

Hate Speech Detection Hate speech detection is the task of identifying abusive language that is specifically directed towards a particular group.To study biases in hate speech detection, many existing works have formulated different datasets and bias metrics. Davidson et al. (2017) and Founta et al. (2018) annotate Twitter datasets for hate speech detection. Blodgett et al. (2016) provide

a corpus of demographically-aligned text with geolocated messages based on Twitter. Sap et al. (2019); Xia et al. (2020) use those datasets to show racial biases through a higher false positive rate for AAE, while Davidson et al. (2019) use the dataset of Blodgett et al. (2016) for racial bias evaluation by comparing probabilities of tweets from different social groups being predicted as hate speech. Davani et al. (2020) collect a dataset of comments from the Gab platform, but analyze biases by comparing a language model's log likelihood differences for constructed counterfactuals. Goldfarb-Tarrant et al. (2020) add gender labels to the dataset from Founta et al. (2018) to analyze gender bias in hate speech detection, and further use Basile et al. (2019)'s multilingual dataset to measure hate speech targeted at women and immigrants in English and Spanish. Similar to toxicity detection, most of these measures demonstrate the harm of online comments across demographic groups through a comparative score.

Bias Analyses without Complete Bias Measures There are other task-specific discussions of bias evaluations that do not propose specific bias measures. For the task of common sense inference (incorporating common sense knowledge into model inference), Rashkin et al. (2018) analyze the intents of entities involved in an event, finding gender differences in the intents. For named entity recognition, Mehrabi et al. (2020) discuss how models have different abilities to recognize male and female names as entities. For part-of-speech tagging, Munro and Morrison (2020) and Garimella et al. (2019) find that state-of-the-art parsers perform differently across genders, failing to identify "hers" and "theirs" as pronouns but not "his". In addition, Mehrabi et al. (2021) and Rudinger et al. (2017) demonstrate severe disparities in common sense knowledge and NLI datasets, respectively.

B.2 Natural Language Generation

We briefly describe some datasets and metrics used to evaluate biases in NLG tasks and refer readers to Sheng et al. (2021) for a survey on common bias measures in Natural Language Generation. For autocomplete generation, Sheng et al. (2019) and Huang et al. (2020) both curate sets of prompts containing different demographic groups to prompt for inequalities in generated text. For the similar task of dialogue generation, Liu et al. (2020a) construct a Twitter-based dataset with parallel context pairs between different groups, and Liu et al. (2020b) rely on extracted conversation and movie datasets to evaluate gender biases. Both works use various metrics such as sentiment, offensiveness, and the occurrence of specific words. For machine translation, the English WinoMT dataset (Stanovsky et al., 2019) is a widely used dataset for quantifying gender biases with bias metrics for translation typically rely on translation accuracy.

C Documenting Bias Measures

C.1 Case Study #1: Documentation for WinoBias (Zhao et al., 2018)

1. *Motivation*
   - **What is the stated definition of bias? How does this definition align with normative definitions of harm?** The paper defines gender bias in coreference resolution as the instance when a system associates pronouns to occupations that are dominated by the pronoun's associated gender more accurately than occupations *not* dominated by that gender. While gendered associations with occupations are an instance of gender bias, such a definition does not capture gender bias in its entirety. The metric is defined to measure occupational perception of different genders, which is associated with representational harms.
   - **What language and culture is the bias and measure most relevant to?** English language in the United States
   - **If a demographic attribute is split into groups for measurement of bias, how many groups have been considered?** Gender binary (male and female) is considered in this measure.
   - **What is the source of bias that is measured?** The paper highlights two sources of gender bias: *training data bias* and *resource bias*. Training data used for coreference resolution systems are noted to have severe gender imbalance (over 80% of entities headed by gendered pronouns are male). Pre-trained word embeddings which serve as an auxiliary resource for WinoBias (Zhao et al., 2018) have been shown to contain gender bias as well ("*men*" is closer to "*programmer*" than "*woman*"). The paper also mentions a gender statistics corpus (i.e. *Gender Lists*) as a resource that contains an uneven number of gendered contexts in which a noun phrase is observed.

- **What tasks or applications is this bias measure useful for?** Since coreference resolution serves as an important step for many higher-level natural language understanding such as information extraction, document summarization, and question answering, this bias metric is useful for any of such tasks.

2. *Composition and Collection Process*
   - **Is the bias measure data scraped, generated, or produced some other way?** The data is created by the authors but the occupation list is collected from the U.S. Bureau of Labor Statistics. An advantage of this is that the profession categories come from an objective, rather than a biased, source as it is a government document. A disadvantage of this is that it is not comprehensive, and it is generated with the narrow view of only the United States.
   - **What are the limitations associated with method of data curation? How generalizable is this dataset?** The data is limited because the occupations are collected from one source, and the source is specific to the United States. We expect that occupation titles and categories vary among different countries. Additionally, it is important to note that the statistics are constantly changing, and although the website that the data updates regularly, the dataset is static. This limits the relevance of the dataset as the world around it changes.
   - **Is the dataset at risk of causing harm through the particular selection of proxy attributes to represent demographic groups?** Possibly—the dataset uses a limited set of occupations (curated from US-specific resources) and binary pronouns to represent different gender groups.

3. *Bias Metrics*
   - **How is the bias metric defined?** It is defined as the absolute score difference between pro-stereotyped and anti-stereotyped conditions, where for pro-stereotypical condition, the gender pronoun is linked with the dominated profession and for anti-stereotypical vice versa.
   - **Is it an absolute or relative evaluation?** As it measures the bias through the difference between pro-stereotyped and anti-stereotyped conditions, it belongs to relative evaluation. Using a relative evaluation allows more flexibility for different models.
   - **Are there alternate or existing metrics this metric can or should be used with?** WinoBias (Zhao et al., 2018) adapts the absolute difference of F1 to evaluate the gender bias. Since F1 score is a general metric to compare model performance, similar to the difference, the ratio could also be used to so disparity between to sets.
   - **Are there other existing datasets or metrics to evaluate bias for the same task?** Yes, for coreference resolution task, there are also Gendered Ambiguous Pronouns (GAP) (Webster et al., 2018) measuring the disparity incorrectly solving pronoun-name relationships for male and female genders, MAP (Cao and Daumé III, 2020) (built on GAP beyond binary genders) and Winogender (Rudinger et al., 2018) which also measures the relationship between gendered pronouns and occupations.
   - **Can the metric imply an absolute absence of bias in a specific task or model?** No, as discussed before, this metric only focuses on entities with 40 occupations in limited sentence templates. Even if the absolute difference doesn't show much inequalities, there could still be biases in the model.

C.2 Case Study #2: Documentation for Regard (Sheng et al., 2019)

1. *Motivation*
   - **What is the stated definition of bias? How does this definition align with normative definitions of harm?** The authors do not provide an explicit definition of bias, but define bias in terms of the metric of *regard* (i.e., social perception) towards a demographic, which can be negative, neutral or positive. Since this metric is defined to measure social perception, it is aligned with definitions of representational harms, e.g., negative stereotypes, denigrations.
   - **What is the source of bias that is measured?** It is difficult to pinpoint the exact sources of biases from the probing experiments run by Sheng et al. (2019) on GPT-2 and the 1 Billion Word Language Model, though we can form hypotheses. While the One Billion Word Benchmark dataset is publicly available for analysis, the exact dataset used to train GPT-2

can probably only be approximated at best. However, we know that GPT-2 was trained on Web data, including from Web sources such as Reddit, which the authors mention as a likely source of biases. The 1 Billion Word Language Model was trained on news data, and Sheng et al. (2019) find less biased results from this model. There could also be non-data related biases (e.g., depending on features in the model architecture and training procedure), though more studies need to be done here.

- **What tasks or applications is this bias measure useful for?** The metric of regard is useful for applications for continuation generation tasks (Sheng et al., 2021), e.g., when a system takes an input prompt and generates text in a mostly unconstrained manner. In other words, this metric could also be useful for dialogue generation, chat bots, virtual assistants, and creative generation applications, in addition to language models.

2. *Composition and Collection Process*
   - **Is the bias measure data scraped, generated, or produced some other way?** The data used as input prompts to probe for biases are generated from templates. For example, *"XYZ worked as"*, *"XYZ earned money by"*, etc. These templates allow for a controlled probing of inequalities in specific contexts related to occupations and respect. The disadvantages are that templates can be time-consuming to manually construct (Sheng et al. (2019) only use 10 templates) and may not be representative or comprehensive of all the ways that similar content could be phrased. Additionally, the templates could be biased towards the syntactic and semantic inclinations of the template creators, which may or may not align with those the model is used to seeing.
   - **What are the limitations associated with method of data curation? How generalizable is this dataset?** These templates are generalizeable to other demographic surface forms not mentioned in original work. Although conceptually these templates can be extended to probe biases in other contexts (e.g., contexts likely to lead to negative religious or ethnic stereotypes), manually creating these contexts is slow and likely not comprehensive. While these templates could also be translated to other languages, relying on automatic translations could result in unnatural phrasings, while manual translations are more time-consuming.

3. *Bias Metrics*
   - **How is the bias metric defined?** Sheng et al. (2019) define the metric of regard (social perception) towards a demographic group. Possible values are negative, neutral, or positive.
   - **Is it an absolute or relative evaluation?** The authors have formatted the comparison of regard scores across demographics as a relative evaluation. Using a relative evaluation allows more flexibility for different analyses.
   - **Are there alternate or existing metrics this metric can or should be used with?** Sheng et al. (2019) show in their study (Table 5) that the metrics of sentiment and regard can be well-correlated for some types of prompts yet greatly differ for other types. They conclude that it could be useful to report both sentiment and regard.
   - **Are there other existing datasets or metrics to evaluate bias for the same task?** At the time of publication, there were perhaps limited proposed alternatives for evaluating biases from language models, though there are now other options. Huang et al. (2020) present 730 manually curated templates to probe for sentiment differences across countries, occupations, and genders in language models. There are also other bias measures for language models that rely on sentiment (Groenwold et al., 2020; Shwartz et al., 2020). Both Sheng et al. (2019) and Huang et al. (2020) construct manual prompts to test for biases towards demographics mentioned in the input. Additionally, Groenwold et al. (2020) evaluate for similar biases in language models towards *people who produce the text* (Sheng et al., 2021). Combining all these bias measures would provide a more comprehensive analysis.
   - **Can the metric imply an absolute absence of bias in a specific task or model?** No, as discussed in earlier answers, the limited templates (both in number and in syntactic/semantic diversity) mean that even if the regard scores do not show inequalities, there could still be biases in the model. Also, since the authors use a regard classifier to feasibly

automatically label a large number of samples, there could also be biases from the classifier itself. Even human evaluations of regard could be influenced by human biases.

Task	Demographic Dimension	Bias Measure	Harms Evaluated
Hate Speech Detection	**Gender** through identity terms	Davani et al. (2020)	**Disparagement, QoS, Stereo.**
	Gender through stereotypes	Founta et al. (2018) + Goldfarb-Tarrant et al. (2020) Basile et al. (2019) + Goldfarb-Tarrant et al. (2020)	**Disparagement** **Dehumanization, Disparagement**
	Migrants through identity terms	Davani et al. (2020)	**Disparagement, QoS, Stereo.**
	Migrants through identity terms	Basile et al. (2019) + Goldfarb-Tarrant et al. (2020)	**Dehumanization, Disparagement** through pleasantness terms
	Political Ideo. through identity terms	Davani et al. (2020)	**Disparagement, QoS, Stereo.**
	Race through dialect	[Blodgett et al. (2016), Davidson et al. (2017), Founta et al. (2018), Preoţiuc-Pietro and Ungar (2018)] + Sap et al. (2019) [Blodgett et al. (2016), Davidson et al. (2017), Founta et al. (2018)] + Xia et al. (2020) [Waseem and Hovy (2016), Waseem (2016), Davidson et al. (2017), Founta et al. (2018), Golbeck et al. (2017), Blodgett et al. (2016)] + Davidson et al. (2019)	**Disparagement, Erasure, QoS** **Disparagement, Erasure** **Disparagement, Erasure, QoS**
	Race through identity terms	Davani et al. (2020) Kennedy et al. (2020)	**Disparagement, QoS, Stereo.** **Dehumanization, Disparagement**
	Religion through identity terms	Davani et al. (2020)	**Disparagement, QoS, Stereo.**
	Sexual Orient. through identity terms	Davani et al. (2020)	**Disparagement, QoS, Stereo.**
MLM Predictions	**Age** through identity terms	Nangia et al. (2020) Neveol et al. (2022)	**Stereo.** **Stereo.**
	Appearance through identity terms	Nangia et al. (2020) Neveol et al. (2022)	**Stereo.** **Stereo.**
	Disability through identity terms	Nangia et al. (2020) Neveol et al. (2022)	**Stereo.** **Stereo.**
	Gender through identity terms	Nangia et al. (2020) Nadeem et al. (2021) Neveol et al. (2022)	**Stereo.** **Stereo.** **Stereo.**
	Nationality through identity terms	Nangia et al. (2020) Neveol et al. (2022)	**Stereo.** **Stereo.**
	Race through identity terms	Nangia et al. (2020) Nadeem et al. (2021) Neveol et al. (2022)	**Stereo.** **Stereo.** **Stereo.**
	Religion through identity terms	Nangia et al. (2020) Nadeem et al. (2021) Neveol et al. (2022)	**Stereo.** **Stereo.** **Stereo.**
	Sexual Orient. through identity terms	Nangia et al. (2020) Neveol et al. (2022)	**Stereo.** **Stereo.**
	Socioeconomic through identity terms	Nangia et al. (2020) Nadeem et al. (2021) Neveol et al. (2022)	**Stereo.** **Stereo.** **Stereo.**
Autocomplete Generation	**Gender** through identity terms	Sheng et al. (2019) Huang et al. (2020) Dhamala et al. (2021)	**Disparagement, Stereo.** **Erasure, Stereo.** **Disparagement, Stereo.**
	Gender through occupations	Alnegheimish et al. (2022)	**Erasure, Stereo.**
	Race through identity terms	Sheng et al. (2019) Dhamala et al. (2021)	**Disparagement, Stereo.** **Disparagement, Stereo.**
	Race through dialect	Groenwold et al. (2020)	**Erasure, Stereo.**
	Sexuality through identity terms	Sheng et al. (2019)	**Disparagement, Stereo.**
	Country through identity terms	Huang et al. (2020)	**Erasure, Stereo.**
	Occupation through identity terms	Huang et al. (2020) Dhamala et al. (2021)	**Erasure, Stereo.** **Disparagement, Stereo.**
	Religion through identity terms	Dhamala et al. (2021)	**Disparagement, Stereo.**
	Political Ideo. through identity terms	Dhamala et al. (2021)	**Disparagement, Stereo.**
Dialogue Generation	**Gender** through identity terms	Liu et al. (2020a,b) Dinan et al. (2020)	**Disparagement, Stereo.** **Dehumanization, Erasure, Stereo.**
	Race through identity terms	Liu et al. (2020a)	**Disparagement, Stereo.**
Translation	**Gender** through occupations	Stanovsky et al. (2019)	**Erasure, QoS, Stereo.**
	Gender through identity terms **Nationality** through identity terms **Race** through identity terms	Wang et al. (2022) Wang et al. (2022) Wang et al. (2022)	**Erasure, QoS, Stereo.** **Erasure, QoS, Stereo.** **Erasure, QoS, Stereo.**
Text Re-writing	**Gender** through inflections	Habash et al. (2019) Zmigrod et al. (2019)	**Erasure, Stereo.** **Erasure, Stereo.**

Table 2: Existing bias measures (pt. 2) by tasks and demographics. '+' means that one work built a bias metric (after '+') on top of a dataset from another (before '+'). Brackets group datasets that were all used by a metric.