Language (Technology) is Power: A Critical Survey of "Bias" in NLP

Su Lin Blodgett
College of Information and Computer Sciences
University of Massachusetts Amherst
`blodgett@cs.umass.edu`

Solon Barocas
Microsoft Research
Cornell University
`solon@microsoft.com`

Hal Daumé III
Microsoft Research
University of Maryland
`me@hal3.name`

Hanna Wallach
Microsoft Research
`wallach@microsoft.com`

Abstract

We survey 146 papers analyzing "bias" in NLP systems, finding that their motivations are often vague, inconsistent, and lacking in normative reasoning, despite the fact that analyzing "bias" is an inherently normative process. We further find that these papers' proposed quantitative techniques for measuring or mitigating "bias" are poorly matched to their motivations and do not engage with the relevant literature outside of NLP. Based on these findings, we describe the beginnings of a path forward by proposing three recommendations that should guide work analyzing "bias" in NLP systems. These recommendations rest on a greater recognition of the relationships between language and social hierarchies, encouraging researchers and practitioners to articulate their conceptualizations of "bias"—i.e., what kinds of system behaviors are harmful, in what ways, to whom, and why, as well as the normative reasoning underlying these statements—and to center work around the lived experiences of members of communities affected by NLP systems, while interrogating and reimagining the power relations between technologists and such communities.

1 Introduction

A large body of work analyzing "bias" in natural language processing (NLP) systems has emerged in recent years, including work on "bias" in embedding spaces (e.g., Bolukbasi et al., 2016a; Caliskan et al., 2017; Gonen and Goldberg, 2019; May et al., 2019) as well as work on "bias" in systems developed for a breadth of tasks including language modeling (Lu et al., 2018; Bordia and Bowman,

2019), coreference resolution (Rudinger et al., 2018; Zhao et al., 2018a), machine translation (Vanmassenhove et al., 2018; Stanovsky et al., 2019), sentiment analysis (Kiritchenko and Mohammad, 2018), and hate speech/toxicity detection (e.g., Park et al., 2018; Dixon et al., 2018), among others.

Although these papers have laid vital groundwork by illustrating some of the ways that NLP systems can be harmful, the majority of them fail to engage critically with what constitutes "bias" in the first place. Despite the fact that analyzing "bias" is an inherently normative process—in which some system behaviors are deemed good and others harmful—papers on "bias" in NLP systems are rife with unstated assumptions about what kinds of system behaviors are harmful, in what ways, to whom, and why. Indeed, the term "bias" (or "gender bias" or "racial bias") is used to describe a wide range of system behaviors, even though they may be harmful in different ways, to different groups, or for different reasons. Even papers analyzing "bias" in NLP systems developed for the same task often conceptualize it differently.

For example, the following system behaviors are all understood to be self-evident statements of "racial bias": (a) embedding spaces in which embeddings for names associated with African Americans are closer (compared to names associated with European Americans) to unpleasant words than pleasant words (Caliskan et al., 2017); (b) sentiment analysis systems yielding different intensity scores for sentences containing names associated with African Americans and sentences containing names associated with European Americans (Kiritchenko and Mohammad, 2018); and (c) toxicity

detection systems scoring tweets containing features associated with African-American English as more offensive than tweets without these features (Davidson et al., 2019; Sap et al., 2019). Moreover, some of these papers focus on "racial bias" expressed in written text, while others focus on "racial bias" against authors. This use of imprecise terminology obscures these important differences.

We survey 146 papers analyzing "bias" in NLP systems, finding that their motivations are often vague and inconsistent. Many lack any normative reasoning for why the system behaviors that are described as "bias" are harmful, in what ways, and to whom. Moreover, the vast majority of these papers do not engage with the relevant literature outside of NLP to ground normative concerns when proposing quantitative techniques for measuring or mitigating "bias." As a result, we find that many of these techniques are poorly matched to their motivations, and are not comparable to one another.

We then describe the beginnings of a path forward by proposing three recommendations that should guide work analyzing "bias" in NLP systems. We argue that such work should examine the relationships between language and social hierarchies; we call on researchers and practitioners conducting such work to articulate their conceptualizations of "bias" in order to enable conversations about what kinds of system behaviors are harmful, in what ways, to whom, and why; and we recommend deeper engagements between technologists and communities affected by NLP systems. We also provide several concrete research questions that are implied by each of our recommendations.

2   Method

Our survey includes all papers known to us analyzing "bias" in NLP systems—146 papers in total. We omitted papers about speech, restricting our survey to papers about written text only. To identify the 146 papers, we first searched the ACL Anthology[1] for all papers with the keywords "bias" or "fairness" that were made available prior to May 2020. We retained all papers about social "bias," and discarded all papers about other definitions of the keywords (e.g., hypothesis-only bias, inductive bias, media bias). We also discarded all papers using "bias" in NLP systems to measure social "bias" in text or the real world (e.g., Garg et al., 2018).

To ensure that we did not exclude any relevant

NLP task	Papers
Embeddings (type-level or contextualized)	54
Coreference resolution	20
Language modeling or dialogue generation	17
Hate-speech detection	17
Sentiment analysis	15
Machine translation	8
Tagging or parsing	5
Surveys, frameworks, and meta-analyses	20
Other	22

Table 1: The NLP tasks covered by the 146 papers.

papers without the keywords "bias" or "fairness," we also traversed the citation graph of our initial set of papers, retaining any papers analyzing "bias" in NLP systems that are cited by or cite the papers in our initial set. Finally, we manually inspected any papers analyzing "bias" in NLP systems from leading machine learning, human–computer interaction, and web conferences and workshops, such as ICML, NeurIPS, AIES, FAccT, CHI, and WWW, along with any relevant papers that were made available in the "Computation and Language" and "Computers and Society" categories on arXiv prior to May 2020, but found that they had already been identified via our traversal of the citation graph. We provide a list of all 146 papers in the appendix. In Table 1, we provide a breakdown of the NLP tasks covered by the papers. We note that counts do not sum to 146, because some papers cover multiple tasks. For example, a paper might test the efficacy of a technique for mitigating "bias" in embedding spaces in the context of sentiment analysis.

Once identified, we then read each of the 146 papers with the goal of categorizing their motivations and their proposed quantitative techniques for measuring or mitigating "bias." We used a previously developed taxonomy of harms for this categorization, which differentiates between so-called *allocational* and *representational* harms (Barocas et al., 2017; Crawford, 2017). Allocational harms arise when an automated system allocates resources (e.g., credit) or opportunities (e.g., jobs) unfairly to different social groups; representational harms arise when a system (e.g., a search engine) represents some social groups in a less favorable light than others, demeans them, or fails to recognize their existence altogether. Adapting and extending this taxonomy, we categorized the 146 papers' motivations and techniques into the following categories:

▷ *Allocational harms.*

	Papers	
Category	Motivation	Technique
Allocational harms	30	4
Stereotyping	50	58
Other representational harms	52	43
Questionable correlations	47	42
Vague/unstated	23	0
Surveys, frameworks, and meta-analyses	20	20

Table 2: The categories into which the 146 papers fall.

▷ *Representational harms:*[2]

    ▷ *Stereotyping* that propagates negative generalizations about particular social groups.

    ▷ Differences in *system performance* for different social groups, language that *misrepresents* the distribution of different social groups in the population, or language that is *denigrating* to particular social groups.

▷ *Questionable correlations* between system behavior and features of language that are typically associated with particular social groups.

▷ *Vague descriptions* of "bias" (or "gender bias" or "racial bias") or *no description* at all.

▷ *Surveys, frameworks, and meta-analyses*.

In Table 2 we provide counts for each of the six categories listed above. (We also provide a list of the papers that fall into each category in the appendix.) Again, we note that the counts do not sum to 146, because some papers state multiple motivations, propose multiple techniques, or propose a single technique for measuring or mitigating multiple harms. Table 3, which is in the appendix, contains examples of the papers' motivations and techniques across a range of different NLP tasks.

3 Findings

Categorizing the 146 papers' motivations and proposed quantitative techniques for measuring or mitigating "bias" into the six categories listed above enabled us to identify several commonalities, which we present below, along with illustrative quotes.

---

[2]We grouped several types of representational harms into two categories to reflect that the main point of differentiation between the 146 papers' motivations and proposed quantitative techniques for measuring or mitigating "bias" is whether or not they focus on stereotyping. Among the papers that do not focus on stereotyping, we found that most lack sufficiently clear motivations and techniques to reliably categorize them further.

3.1 Motivations

Papers state a wide range of motivations, multiple motivations, vague motivations, and sometimes no motivations at all. We found that the papers' motivations span all six categories, with several papers falling into each one. Appropriately, papers that provide surveys or frameworks for analyzing "bias" in NLP systems often state multiple motivations (e.g., Hovy and Spruit, 2016; Bender, 2019; Sun et al., 2019; Rozado, 2020; Shah et al., 2020). However, as the examples in Table 3 (in the appendix) illustrate, many other papers (33%) do so as well. Some papers (16%) state only vague motivations or no motivations at all. For example,

> *"[N]o human should be discriminated on the basis of demographic attributes by an NLP system."*
> —Kaneko and Bollegala (2019)

> *"[P]rominent word embeddings [...] encode systematic biases against women and black people [...] implicating many NLP systems in scaling up social injustice."* —May et al. (2019)

These examples leave unstated what it might mean for an NLP system to "discriminate," what constitutes "systematic biases," or how NLP systems contribute to "social injustice" (itself undefined).

Papers' motivations sometimes include no normative reasoning. We found that some papers (32%) are not motivated by any apparent normative concerns, often focusing instead on concerns about system performance. For example, the first quote below includes normative reasoning—namely that models should not use demographic information to make predictions—while the other focuses on learned correlations impairing system performance.

> *"In [text classification], models are expected to make predictions with the semantic information rather than with the demographic group identity information (e.g., 'gay', 'black') contained in the sentences."* —Zhang et al. (2020a)

> *"An over-prevalence of some gendered forms in the training data leads to translations with identifiable errors. Translations are better for sentences involving men and for sentences containing stereotypical gender roles."*
> —Saunders and Byrne (2020)

Even when papers do state clear motivations, they are often unclear about why the system behaviors that are described as "bias" are harmful, in what ways, and to whom. We found that even papers with clear motivations often fail to explain what kinds of system behaviors are harmful, in what ways, to whom, and why. For example,

"Deploying these word embedding algorithms in practice, for example in automated translation systems or as hiring aids, runs the serious risk of perpetuating problematic biases in important societal contexts." —Brunet et al. (2019)

"[I]f the systems show discriminatory behaviors in the interactions, the user experience will be adversely affected." —Liu et al. (2019)

These examples leave unstated what "problematic biases" or non-ideal user experiences might look like, how the system behaviors might result in these things, and who the relevant stakeholders or users might be. In contrast, we find that papers that provide surveys or frameworks for analyzing "bias" in NLP systems often name who is harmed, acknowledging that different social groups may experience these systems differently due to their different relationships with NLP systems or different social positions. For example, Ruane et al. (2019) argue for a "deep understanding of the user groups [sic] characteristics, contexts, and interests" when designing conversational agents.

Papers about NLP systems developed for the same task often conceptualize "bias" differently. Even papers that cover the same NLP task often conceptualize "bias" in ways that differ substantially and are sometimes inconsistent. Rows 3 and 4 of Table 3 (in the appendix) contain machine translation papers with different conceptualizations of "bias," leading to different proposed techniques, while rows 5 and 6 contain papers on "bias" in embedding spaces that state different motivations, but propose techniques for quantifying stereotyping.

Papers' motivations conflate allocational and representational harms. We found that the papers' motivations sometimes (16%) name immediate representational harms, such as stereotyping, alongside more distant allocational harms, which, in the case of stereotyping, are usually imagined as downstream effects of stereotypes on résumé filtering. Many of these papers use the imagined downstream effects to justify focusing on particular system behaviors, even when the downstream effects are not measured. Papers on "bias" in embedding spaces are especially likely to do this because embeddings are often used as input to other systems:

"However, none of these papers [on embeddings] have recognized how blatantly sexist the embeddings are and hence risk introducing biases of various types into real-world systems." —Bolukbasi et al. (2016a)

"It is essential to quantify and mitigate gender bias in these embeddings to avoid them from affecting downstream applications." —Zhou et al. (2019)

In contrast, papers that provide surveys or frameworks for analyzing "bias" in NLP systems treat representational harms as harmful in their own right. For example, Mayfield et al. (2019) and Ruane et al. (2019) cite the harmful reproduction of dominant linguistic norms by NLP systems (a point to which we return in section 4), while Bender (2019) outlines a range of harms, including seeing stereotypes in search results and being made invisible to search engines due to language practices.

3.2 Techniques

Papers' techniques are not well grounded in the relevant literature outside of NLP. Perhaps unsurprisingly given that the papers' motivations are often vague, inconsistent, and lacking in normative reasoning, we also found that the papers' proposed quantitative techniques for measuring or mitigating "bias" do not effectively engage with the relevant literature outside of NLP. Papers on stereotyping are a notable exception: the Word Embedding Association Test (Caliskan et al., 2017) draws on the Implicit Association Test (Greenwald et al., 1998) from the social psychology literature, while several techniques operationalize the well-studied "Angry Black Woman" stereotype (Kiritchenko and Mohammad, 2018; May et al., 2019; Tan and Celis, 2019) and the "double bind" faced by women (May et al., 2019; Tan and Celis, 2019), in which women who succeed at stereotypically male tasks are perceived to be less likable than similarly successful men (Heilman et al., 2004). Tan and Celis (2019) also examine the compounding effects of race and gender, drawing on Black feminist scholarship on intersectionality (Crenshaw, 1989).

Papers' techniques are poorly matched to their motivations. We found that although 21% of the papers include allocational harms in their motivations, only four papers actually propose techniques for measuring or mitigating allocational harms.

Papers focus on a narrow range of potential sources of "bias." We found that nearly all of the papers focus on system predictions as the potential sources of "bias," with many additionally focusing on "bias" in datasets (e.g., differences in the number of gendered pronouns in the training data (Zhao et al., 2019)). Most papers do not interrogate

the normative implications of other decisions made during the development and deployment lifecycle—perhaps unsurprising given that their motivations sometimes include no normative reasoning. A few papers are exceptions, illustrating the impacts of task definitions, annotation guidelines, and evaluation metrics: Cao and Daumé (2019) study how folk conceptions of gender (Keyes, 2018) are reproduced in coreference resolution systems that assume a strict gender dichotomy, thereby maintaining cisnormativity; Sap et al. (2019) focus on the effect of priming annotators with information about possible dialectal differences when asking them to apply toxicity labels to sample tweets, finding that annotators who are primed are significantly less likely to label tweets containing features associated with African-American English as offensive.

4   A path forward

We now describe how researchers and practitioners conducting work analyzing "bias" in NLP systems might avoid the pitfalls presented in the previous section—the beginnings of a path forward. We propose three recommendations that should guide such work, and, for each, provide several concrete research questions. We emphasize that these questions are not comprehensive, and are intended to generate further questions and lines of engagement. Our three recommendations are as follows:

(R1) Ground work analyzing "bias" in NLP systems in the relevant literature outside of NLP that explores the relationships between language and social hierarchies. Treat representational harms as harmful in their own right.

(R2) Provide explicit statements of why the system behaviors that are described as "bias" are harmful, in what ways, and to whom. Be forthright about the normative reasoning (Green, 2019) underlying these statements.

(R3) Examine language use in practice by engaging with the lived experiences of members of communities affected by NLP systems. Interrogate and reimagine the power relations between technologists and such communities.

4.1   Language and social hierarchies

Turning first to **(R1)**, we argue that work analyzing "bias" in NLP systems will paint a much fuller picture if it engages with the relevant literature outside of NLP that explores the relationships between

language and social hierarchies. Many disciplines, including sociolinguistics, linguistic anthropology, sociology, and social psychology, study how language takes on social meaning and the role that language plays in maintaining social hierarchies. For example, language is the means through which social groups are labeled and one way that beliefs about social groups are transmitted (e.g., Maass, 1999; Beukeboom and Burgers, 2019). Group labels can serve as the basis of stereotypes and thus reinforce social inequalities: "[T]he label content functions to identify a given category of people, and thereby conveys category boundaries and a position in a hierarchical taxonomy" (Beukeboom and Burgers, 2019). Similarly, "controlling images," such as stereotypes of Black women, which are linguistically and visually transmitted through literature, news media, television, and so forth, provide "ideological justification" for their continued oppression (Collins, 2000, Chapter 4).

As a result, many groups have sought to bring about social changes through changes in language, disrupting patterns of oppression and marginalization via so-called "gender-fair" language (Sczesny et al., 2016; Menegatti and Rubini, 2017), language that is more inclusive to people with disabilities (ADA, 2018), and language that is less dehumanizing (e.g., abandoning the use of the term "illegal" in everyday discourse on immigration in the U.S. (Rosa, 2019)). The fact that group labels are so contested is evidence of how deeply intertwined language and social hierarchies are. Taking "gender-fair" language as an example, the hope is that reducing asymmetries in language about women and men will reduce asymmetries in their social standing. Meanwhile, struggles over language use often arise from dominant social groups' desire to "control both material and symbolic resources"—i.e., "the right to decide what words will mean and to control those meanings"—as was the case in some white speakers' insistence on using offensive place names against the objections of Indigenous speakers (Hill, 2008, Chapter 3).

Sociolinguists and linguistic anthropologists have also examined language attitudes and language ideologies, or people's metalinguistic beliefs about language: Which language varieties or practices are taken as standard, ordinary, or unmarked? Which are considered correct, prestigious, or appropriate for public use, and which are considered incorrect, uneducated, or offensive (e.g., Campbell-

Kibler, 2009; Preston, 2009; Loudermilk, 2015; Lanehart and Malik, 2018)? Which are rendered invisible (Roche, 2019)?[3] Language ideologies play a vital role in reinforcing and justifying social hierarchies because beliefs about language varieties or practices often translate into beliefs about their speakers (e.g. Alim et al., 2016; Rosa and Flores, 2017; Craft et al., 2020). For example, in the U.S., the portrayal of non-white speakers' language varieties and practices as linguistically deficient helped to justify violent European colonialism, and today continues to justify enduring racial hierarchies by maintaining views of non-white speakers as lacking the language "required for complex thinking processes and successful engagement in the global economy" (Rosa and Flores, 2017).

Recognizing the role that language plays in maintaining social hierarchies is critical to the future of work analyzing "bias" in NLP systems. First, it helps to explain why representational harms are harmful in their own right. Second, the complexity of the relationships between language and social hierarchies illustrates why studying "bias" in NLP systems is so challenging, suggesting that researchers and practitioners will need to move beyond existing algorithmic fairness techniques. We argue that work must be grounded in the relevant literature outside of NLP that examines the relationships between language and social hierarchies; without this grounding, researchers and practitioners risk measuring or mitigating only what is convenient to measure or mitigate, rather than what is most normatively concerning.

More specifically, we recommend that work analyzing "bias" in NLP systems be reoriented around the following question: How are social hierarchies, language ideologies, and NLP systems coproduced? This question mirrors Benjamin's (2020) call to examine how "race and technology are coproduced"—i.e., how racial hierarchies, and the ideologies and discourses that maintain them, create and are re-created by technology. We recommend that researchers and practitioners similarly ask how existing social hierarchies and language ideologies drive the development and deployment of NLP systems, and how these systems therefore reproduce these hierarchies and ideologies. As a starting point for reorienting work analyzing "bias" in NLP systems around this question, we

_____

[3]Language ideologies encompass much more than this; see, e.g., Lippi-Green (2012), Alim et al. (2016), Rosa and Flores (2017), Rosa and Burdick (2017), and Charity Hudley (2017).

provide the following concrete research questions:

▷ How do social hierarchies and language ideologies influence the decisions made during the development and deployment lifecycle? What kinds of NLP systems do these decisions result in, and what kinds do they foreclose?

  ◇ General assumptions: To which linguistic norms do NLP systems adhere (Bender, 2019; Ruane et al., 2019)? Which language practices are implicitly assumed to be standard, ordinary, correct, or appropriate?

  ◇ Task definition: For which speakers are NLP systems (and NLP resources) developed? (See Joshi et al. (2020) for a discussion.) How do task definitions discretize the world? For example, how are social groups delineated when defining demographic attribute prediction tasks (e.g., Koppel et al., 2002; Rosenthal and McKeown, 2011; Nguyen et al., 2013)? What about languages in native language prediction tasks (Tetreault et al., 2013)?

  ◇ Data: How are datasets collected, preprocessed, and labeled or annotated? What are the impacts of annotation guidelines, annotator assumptions and perceptions (Olteanu et al., 2019; Sap et al., 2019; Geiger et al., 2020), and annotation aggregation processes (Pavlick and Kwiatkowski, 2019)?

  ◇ Evaluation: How are NLP systems evaluated? What are the impacts of evaluation metrics (Olteanu et al., 2017)? Are any non-quantitative evaluations performed?

▷ How do NLP systems reproduce or transform language ideologies? Which language varieties or practices come to be deemed good or bad? Might "good" language simply mean language that is easily handled by existing NLP systems? For example, linguistic phenomena arising from many language practices (Eisenstein, 2013) are described as "noisy text" and often viewed as a target for "normalization." How do the language ideologies that are reproduced by NLP systems maintain social hierarchies?

▷ Which representational harms are being measured or mitigated? Are these the most normatively concerning harms, or merely those that are well handled by existing algorithmic fairness techniques? Are there other representational harms that might be analyzed?

4.2 Conceptualizations of "bias"

Turning now to **(R2)**, we argue that work analyzing "bias" in NLP systems should provide explicit statements of why the system behaviors that are described as "bias" are harmful, in what ways, and to whom, as well as the normative reasoning underlying these statements. In other words, researchers and practitioners should articulate their conceptualizations of "bias." As we described above, papers often contain descriptions of system behaviors that are understood to be self-evident statements of "bias." This use of imprecise terminology has led to papers all claiming to analyze "bias" in NLP systems, sometimes even in systems developed for the same task, but with different or even inconsistent conceptualizations of "bias," and no explanations for these differences.

Yet analyzing "bias" is an inherently normative process—in which some system behaviors are deemed good and others harmful—even if assumptions about what kinds of system behaviors are harmful, in what ways, for whom, and why are not stated. We therefore echo calls by Bardzell and Bardzell (2011), Keyes et al. (2019), and Green (2019) for researchers and practitioners to make their normative reasoning explicit by articulating the social values that underpin their decisions to deem some system behaviors as harmful, no matter how obvious such values appear to be. We further argue that this reasoning should take into account the relationships between language and social hierarchies that we described above. First, these relationships provide a foundation from which to approach the normative reasoning that we recommend making explicit. For example, some system behaviors might be harmful precisely because they maintain social hierarchies. Second, if work analyzing "bias" in NLP systems is reoriented to understand how social hierarchies, language ideologies, and NLP systems are coproduced, then this work will be incomplete if we fail to account for the ways that social hierarchies and language ideologies determine what we mean by "bias" in the first place. As a starting point, we therefore provide the following concrete research questions:

▷ What kinds of system behaviors are described as "bias"? What are their potential sources (e.g., general assumptions, task definition, data)?

▷ In what ways are these system behaviors harmful, to whom are they harmful, and why?

▷ What are the social values (obvious or not) that

underpin this conceptualization of "bias?"

4.3 Language use in practice

Finally, we turn to **(R3)**. Our perspective, which rests on a greater recognition of the relationships between language and social hierarchies, suggests several directions for examining language use in practice. Here, we focus on two. First, because language is necessarily situated, and because different social groups have different lived experiences due to their different social positions (Hanna et al., 2020)—particularly groups at the intersections of multiple axes of oppression—we recommend that researchers and practitioners center work analyzing "bias" in NLP systems around the lived experiences of members of communities affected by these systems. Second, we recommend that the power relations between technologists and such communities be interrogated and reimagined. Researchers have pointed out that algorithmic fairness techniques, by proposing incremental technical mitigations—e.g., collecting new datasets or training better models—maintain these power relations by (a) assuming that automated systems should continue to exist, rather than asking whether they should be built at all, and (b) keeping development and deployment decisions in the hands of technologists (Bennett and Keyes, 2019; Cifor et al., 2019; Green, 2019; Katell et al., 2020).

There are many disciplines for researchers and practitioners to draw on when pursuing these directions. For example, in human–computer interaction, Hamidi et al. (2018) study transgender people's experiences with automated gender recognition systems in order to uncover how these systems reproduce structures of transgender exclusion by redefining what it means to perform gender "normally." Value-sensitive design provides a framework for accounting for the values of different stakeholders in the design of technology (e.g., Friedman et al., 2006; Friedman and Hendry, 2019; Le Dantec et al., 2009; Yoo et al., 2019), while participatory design seeks to involve stakeholders in the design process itself (Sanders, 2002; Muller, 2007; Simonsen and Robertson, 2013; DiSalvo et al., 2013). Participatory action research in education (Kemmis, 2006) and in language documentation and reclamation (Junker, 2018) is also relevant. In particular, work on language reclamation to support decolonization and tribal sovereignty (Leonard, 2012) and work in sociolinguistics focus-

ing on developing co-equal research relationships with community members and supporting linguistic justice efforts (e.g., Bucholtz et al., 2014, 2016, 2019) provide examples of more emancipatory relationships with communities. Finally, several workshops and events have begun to explore how to empower stakeholders in the development and deployment of technology (Vaccaro et al., 2019; Givens and Morris, 2020; Sassaman et al., 2020)[4] and how to help researchers and practitioners consider when not to build systems at all (Barocas et al., 2020).

As a starting point for engaging with communities affected by NLP systems, we therefore provide the following concrete research questions:

▷ How do communities become aware of NLP systems? Do they resist them, and if so, how?

▷ What additional costs are borne by communities for whom NLP systems do not work well?

▷ Do NLP systems shift power toward oppressive institutions (e.g., by enabling predictions that communities do not want made, linguistically based unfair allocation of resources or opportunities (Rosa and Flores, 2017), surveillance, or censorship), or away from such institutions?

▷ Who is involved in the development and deployment of NLP systems? How do decision-making processes maintain power relations between technologists and communities affected by NLP systems? Can these processes be changed to reimagine these relations?

5   Case study

To illustrate our recommendations, we present a case study covering work on African-American English (AAE).[5] Work analyzing "bias" in the context of AAE has shown that part-of-speech taggers, language identification systems, and dependency parsers all work less well on text containing features associated with AAE than on text without these features (Jørgensen et al., 2015, 2016; Blodgett et al., 2016, 2018), and that toxicity detection systems score tweets containing features associated with AAE as more offensive than tweets without them (Davidson et al., 2019; Sap et al., 2019).

These papers have been critical for highlighting AAE as a language variety for which existing NLP

systems may not work, illustrating their limitations. However, they do not conceptualize "racial bias" in the same way. The first four of these papers simply focus on system performance differences between text containing features associated with AAE and text without these features. In contrast, the last two papers also focus on such system performance differences, but motivate this focus with the following additional reasoning: If tweets containing features associated with AAE are scored as more offensive than tweets without these features, then this might (a) yield negative perceptions of AAE; (b) result in disproportionate removal of tweets containing these features, impeding participation in online platforms and reducing the space available online in which speakers can use AAE freely; and (c) cause AAE speakers to incur additional costs if they have to change their language practices to avoid negative perceptions or tweet removal.

More importantly, none of these papers engage with the literature on AAE, racial hierarchies in the U.S., and raciolinguistic ideologies. By failing to engage with this literature—thereby treating AAE simply as one of many non-Penn Treebank varieties of English or perhaps as another challenging domain—work analyzing "bias" in NLP systems in the context of AAE fails to situate these systems in the world. Who are the speakers of AAE? How are they viewed? We argue that AAE as a language variety cannot be separated from its speakers— primarily Black people in the U.S., who experience systemic anti-Black racism—and the language ideologies that reinforce and justify racial hierarchies.

Even after decades of sociolinguistic efforts to legitimize AAE, it continues to be viewed as "bad" English and its speakers continue to be viewed as linguistically inadequate—a view called the *deficit* perspective (Alim et al., 2016; Rosa and Flores, 2017). This perspective persists despite demonstrations that AAE is rule-bound and grammatical (Mufwene et al., 1998; Green, 2002), in addition to ample evidence of its speakers' linguistic adroitness (e.g., Alim, 2004; Rickford and King, 2016). This perspective belongs to a broader set of raciolinguistic ideologies (Rosa and Flores, 2017), which also produce allocational harms; speakers of AAE are frequently penalized for not adhering to dominant language practices, including in the education system (Alim, 2004; Terry et al., 2010), when seeking housing (Baugh, 2018), and in the judicial system, where their testimony is misunderstood or,

---

[4] Also `https://participatoryml.github.io/`

[5] This language variety has had many different names over the years, but is now generally called African-American English (AAE), African-American Vernacular English (AAVE), or African-American Language (AAL) (Green, 2002; Wolfram and Schilling, 2015; Rickford and King, 2016).

worse yet, disbelieved (Rickford and King, 2016; Jones et al., 2019). These raciolinguistic ideologies position racialized communities as needing linguistic intervention, such as language education programs, in which these and other harms can be reduced if communities accommodate to dominant language practices (Rosa and Flores, 2017).

In the technology industry, speakers of AAE are often not considered consumers who matter. For example, Benjamin (2019) recounts an Apple employee who worked on speech recognition for Siri:

> *"As they worked on different English dialects —Australian, Singaporean, and Indian English — [the employee] asked his boss: 'What about African American English?' To this his boss responded: 'Well, Apple products are for the premium market."'*

The reality, of course, is that speakers of AAE tend not to represent the "premium market" precisely because of institutions and policies that help to maintain racial hierarchies by systematically denying them the opportunities to develop wealth that are available to white Americans (Rothstein, 2017)—an exclusion that is reproduced in technology by countless decisions like the one described above.

Engaging with the literature outlined above situates the system behaviors that are described as "bias," providing a foundation for normative reasoning. Researchers and practitioners should be concerned about "racial bias" in toxicity detection systems not only because performance differences impair system performance, but because they reproduce longstanding injustices of stigmatization and disenfranchisement for speakers of AAE. In re-stigmatizing AAE, they reproduce language ideologies in which AAE is viewed as ungrammatical, uneducated, and offensive. These ideologies, in turn, enable linguistic discrimination and justify enduring racial hierarchies (Rosa and Flores, 2017). Our perspective, which understands racial hierarchies and raciolinguistic ideologies as structural conditions that govern the development and deployment of technology, implies that techniques for measuring or mitigating "bias" in NLP systems will necessarily be incomplete unless they interrogate and dismantle these structural conditions, including the power relations between technologists and racialized communities.

We emphasize that engaging with the literature on AAE, racial hierarchies in the U.S., and raciolinguistic ideologies can generate new lines of engagement. These lines include work on the ways that the decisions made during the development and deployment of NLP systems produce stigmatization and disenfranchisement, and work on AAE use in practice, such as the ways that speakers of AAE interact with NLP systems that were not designed for them. This literature can also help researchers and practitioners address the allocational harms that may be produced by NLP systems, and ensure that even well-intentioned NLP systems do not position racialized communities as needing linguistic intervention or accommodation to dominant language practices. Finally, researchers and practitioners wishing to design better systems can also draw on a growing body of work on anti-racist language pedagogy that challenges the deficit perspective of AAE and other racialized language practices (e.g. Flores and Chaparro, 2018; Baker-Bell, 2019; Martínez and Mejía, 2019), as well as the work that we described in section 4.3 on reimagining the power relations between technologists and communities affected by technology.

6 Conclusion

By surveying 146 papers analyzing "bias" in NLP systems, we found that (a) their motivations are often vague, inconsistent, and lacking in normative reasoning; and (b) their proposed quantitative techniques for measuring or mitigating "bias" are poorly matched to their motivations and do not engage with the relevant literature outside of NLP. To help researchers and practitioners avoid these pitfalls, we proposed three recommendations that should guide work analyzing "bias" in NLP systems, and, for each, provided several concrete research questions. These recommendations rest on a greater recognition of the relationships between language and social hierarchies—a step that we see as paramount to establishing a path forward.

Acknowledgments

References

Artem Abzaliev. 2019. On GAP coreference resolution shared task: insights from the 3rd place solution. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, pages 107–112, Florence, Italy.

ADA. 2018. Guidelines for Writing About People With Disabilities. ADA National Network. https://bit.ly/2KREbkB.

Oshin Agarwal, Funda Durupinar, Norman I. Badler, and Ani Nenkova. 2019. Word embeddings (also) encode human personality stereotypes. In *Proceedings of the Joint Conference on Lexical and Computational Semantics*, pages 205–211, Minneapolis, MN.

H. Samy Alim. 2004. *You Know My Steez: An Ethnographic and Sociolinguistic Study of Styleshifting in a Black American Speech Community*. American Dialect Society.

H. Samy Alim, John R. Rickford, and Arnetha F. Ball, editors. 2016. *Raciolinguistics: How Language Shapes Our Ideas About Race*. Oxford University Press.

Sandeep Attree. 2019. Gendered ambiguous pronouns shared task: Boosting model confidence by evidence pooling. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, Florence, Italy.

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *Proceedings of the International World Wide Web Conference*, pages 49–59, San Francisco, CA.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. Differential Privacy Has Disparate Impact on Model Accuracy. In *Proceedings of the Conference on Neural Information Processing Systems*, Vancouver, Canada.

April Baker-Bell. 2019. Dismantling anti-black linguistic racism in English language arts classrooms: Toward an anti-racist black language pedagogy. *Theory Into Practice*.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 2138–2144, Minneapolis, MN.

Xingce Bao and Qianqian Qiao. 2019. Transfer Learning from Pre-trained BERT for Pronoun Resolution. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, pages 82–88, Florence, Italy.

Shaowen Bardzell and Jeffrey Bardzell. 2011. Towards a Feminist HCI Methodology: Social Science, Feminism, and HCI. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 675–684, Vancouver, Canada.

Solon Barocas, Asia J. Biega, Benjamin Fish, Jędrzej Niklas, and Luke Stark. 2020. When Not to Design, Build, or Deploy. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The Problem With Bias: Allocative Versus Representational Harms in Machine Learning. In *Proceedings of SIGCIS*, Philadelphia, PA.

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

John Baugh. 2018. *Linguistics in Pursuit of Justice*. Cambridge University Press.

Emily M. Bender. 2019. A typology of ethical risks in language technology with an eye towards where transparent documentation can help. Presented at The Future of Artificial Intelligence: Language, Ethics, Technology Workshop. https://bit.ly/2P9t9M6.

Ruha Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. John Wiley & Sons.

Ruha Benjamin. 2020. 2020 Vision: Reimagining the Default Settings of Technology & Society. Keynote at ICLR.

Cynthia L. Bennett and Os Keyes. 2019. What is the Point of Fairness? Disability, AI, and The Complexity of Justice. In *Proceedings of the ASSETS Workshop on AI Fairness for People with Disabilities*, Pittsburgh, PA.

Camiel J. Beukeboom and Christian Burgers. 2019. How Stereotypes Are Shared Through Language: A Review and Introduction of the Social Categories and Stereotypes Communication (SCSC) Framework. *Review of Communication Research*, 7:1–37.

Shruti Bhargava and David Forsyth. 2019. Exposing and Correcting the Gender Bias in Image Captioning Datasets and Models. *arXiv preprint arXiv:1912.00578*.

Jayadev Bhaskaran and Isha Bhallamudi. 2019. Good Secretaries, Bad Truck Drivers? Occupational Gender Stereotypes in Sentiment Analysis. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, pages 62–68, Florence, Italy.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 1119–1130, Austin, TX.

Su Lin Blodgett and Brendan O'Connor. 2017. Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English. In *Proceedings of the Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*, Halifax, Canada.

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. 2018. Twitter Universal Dependency Parsing for African-American and Mainstream American English. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1415–1425, Melbourne, Australia.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016a. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 4349–4357, Barcelona, Spain.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016b. Quantifying and reducing stereotypes in word embeddings. In *Proceedings of the ICML Workshop on #Data4Good: Machine Learning in Social Good Applications*, pages 41–45, New York, NY.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the NAACL Student Research Workshop*, pages 7–15, Minneapolis, MN.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the Origins of Bias in Word Embeddings. In *Proceedings of the International Conference on Machine Learning*, pages 803–811, Long Beach, CA.

Mary Bucholtz, Dolores Inés Casillas, and Jin Sook Lee. 2016. Beyond Empowerment: Accompaniment and Sociolinguistic Justice in a Youth Research Program. In Robert Lawson and Dave Sayers, editors, *Sociolinguistic Research: Application and Impact*, pages 25–44. Routledge.

Mary Bucholtz, Dolores Inés Casillas, and Jin Sook Lee. 2019. California Latinx Youth as Agents of Sociolinguistic Justice. In Netta Avineri, Laura R. Graham, Eric J. Johnson, Robin Conley Riner, and Jonathan Rosa, editors, *Language and Social Justice in Practice*, pages 166–175. Routledge.

Mary Bucholtz, Audrey Lopez, Allina Mojarro, Elena Skapoulli, Chris VanderStouwe, and Shawn Warner-Garcia. 2014. Sociolinguistic Justice in the Schools: Student Researchers as Linguistic Experts. *Language and Linguistics Compass*, 8:144–157.

Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also Snowboard: Overcoming Bias in Captioning Models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 793–811, Munich, Germany.

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

Kathryn Campbell-Kibler. 2009. The nature of sociolinguistic perception. *Language Variation and Change*, 21(1):135–156.

Yang Trista Cao and Hal Daumé, III. 2019. Toward gender-inclusive coreference resolution. *arXiv preprint arXiv:1910.13913*.

Rakesh Chada. 2019. Gendered pronoun resolution using bert and an extractive question answering formulation. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, pages 126–133, Florence, Italy.

Kaytlin Chaloner and Alfredo Maldonado. 2019. Measuring Gender Bias in Word Embedding across Domains and Discovering New Gender Bias Word Categories. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, pages 25–32, Florence, Italy.

Anne H. Charity Hudley. 2017. Language and Racialization. In Ofelia García, Nelson Flores, and Massimiliano Spotti, editors, *The Oxford Handbook of Language and Society*. Oxford University Press.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy.

Shivang Chopra, Ramit Sawhney, Puneet Mathur, and Rajiv Ratn Shah. 2020. Hindi-English Hate Speech Detection: Author Profiling, Debiasing, and Practical Perspectives. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, New York, NY.

Marika Cifor, Patricia Garcia, T.L. Cowan, Jasmine Rault, Tonia Sutherland, Anita Say Chan, Jennifer Rode, Anna Lauren Hoffmann, Niloufar Salehi, and Lisa Nakamura. 2019. Feminist Data Manifest-No. Retrieved from https://www.manifestno.com/.

Patricia Hill Collins. 2000. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Routledge.

Justin T. Craft, Kelly E. Wright, Rachel Elizabeth Weissler, and Robin M. Queen. 2020. Language and Discrimination: Generating Meaning, Perceiving Identities, and Discriminating Outcomes. *Annual Review of Linguistics*, 6(1).

Kate Crawford. 2017. The Trouble with Bias. Keynote at NeurIPS.

Kimberle Crenshaw. 1989. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrmination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum*.

Amanda Cercas Curry and Verena Rieser. 2018. #MeToo: How Conversational Systems Respond to Sexual Harassment. In *Proceedings of the Workshop on Ethics in Natural Language Processing*, pages 7–14, New Orleans, LA.

Karan Dabas, Nishtha Madaan, Gautam Singh, Vijay Arya, Sameep Mehta, and Tanmoy Chakraborty. 2020. Fair Transfer of Multiple Style Attributes in Text. *arXiv preprint arXiv:2001.06693*.

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

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, Atlanta, GA.

Sunipa Dev, Tao Li, Jeff Phillips, and Vivek Srikumar. 2019. On Measuring and Mitigating Biased Inferences of Word Embeddings. *arXiv preprint arXiv:1908.09369*.

Sunipa Dev and Jeff Phillips. 2019. Attenuating Bias in Word Vectors. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 879–887, Naha, Japan.

Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, Montréal, Canada.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2019. Queens are Powerful too: Mitigating Gender Bias in Dialogue Generation. *arXiv preprint arXiv:1911.03842*.

Carl DiSalvo, Andrew Clement, and Volkmar Pipek. 2013. Communities: Participatory Design for, with and by communities. In Jesper Simonsen and Toni Robertson, editors, *Routledge International Handbook of Participatory Design*, pages 182–209. Routledge.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the Conference on Artificial Intelligence, Ethics, and Society (AIES)*, New Orleans, LA.

Jacob Eisenstein. 2013. What to do about bad language on the Internet. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 359–369.

Kawin Ethayarajh. 2020. Is Your Classifier Actually Biased? Measuring Fairness under Uncertainty with Bernstein Bounds. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding Undesirable Word Embedding Assocations. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1696–1705, Florence, Italy.

Joseph Fisher. 2019. Measuring social bias in knowledge graph embeddings. *arXiv preprint arXiv:1912.02761*.

Nelson Flores and Sofia Chaparro. 2018. What counts as language education policy? Developing a materialist Anti-racist approach to language activism. *Language Policy*, 17(3):365–384.

Omar U. Florez. 2019. On the Unintended Social Bias of Training Language Generation Models with Data from Local Media. In *Proceedings of the NeurIPS Workshop on Human-Centric Machine Learning*, Vancouver, Canada.

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

Batya Friedman and David G. Hendry. 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press.

Batya Friedman, Peter H. Kahn Jr., and Alan Borning. 2006. Value Sensitive Design and Information Systems. In Dennis Galletta and Ping Zhang, editors, *Human-Computer Interaction in Management Information Systems: Foundations*, pages 348–372. M.E. Sharpe.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *Proceedings of the National Academy of Sciences*, 115(16).

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the Conference on Artificial Intelligence, Ethics, and Society (AIES)*, Honolulu, HI.

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

Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2020. Towards Understanding Gender Bias in Relation Extraction. In *Proceedings of the Association for Computational Linguistics (ACL)*.

R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage In, Garbage Out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From? In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 325–336.

Oguzhan Gencoglu. 2020. Cyberbullying Detection with Fairness Constraints. *arXiv preprint arXiv:2005.06625*.

Alexandra Reeve Givens and Meredith Ringel Morris. 2020. Centering Disability Perspecives in Algorithmic Fairness, Accountability, and Transparency. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain.

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

Hila Gonen and Kellie Webster. 2020. Automatically Identifying Gender Issues in Machine Translation using Perturbations. *arXiv preprint arXiv:2004.14065*.

Ben Green. 2019. "Good" isn't good enough. In *Proceedings of the AI for Social Good Workshop*, Vancouver, Canada.

Lisa J. Green. 2002. *African American English: A Linguistic Introduction*. Cambridge University Press.

Anthony G. Greenwald, Debbie E. McGhee, and Jordan L.K. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464–1480.

Enoch Opanin Gyamfi, Yunbo Rao, Miao Gou, and Yanhua Shao. 2020. deb2viz: Debiasing gender in word embedding data using subspace visualization. In *Proceedings of the International Conference on Graphics and Image Processing*.

Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. 2018. Gender Recognition or Gender Reductionism? The Social Implications of Automatic Gender Recognition Systems. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, Montréal, Canada.

Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a Critical Race Methodology in Algorithmic Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 501–512, Barcelona, Spain.

Madeline E. Heilman, Aaaron S. Wallen, Daniella Fuchs, and Melinda M. Tamkins. 2004. Penalties for Success: Reactions to Women Who Succeed at Male Gender-Typed Tasks. *Journal of Applied Psychology*, 89(3):416–427.

Jane H. Hill. 2008. *The Everyday Language of White Racism*. Wiley-Blackwell.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. Can You Translate that into Man? Commercial Machine Translation Systems Include Stylistic Biases. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Dirk Hovy and Anders Søgaard. 2015. Tagging Performance Correlates with Author Age. In *Proceedings of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 483–488, Beijing, China.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 591–598, Berlin, Germany.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack W. Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2019. Reducing Sentiment Bias in Language Models via Counterfactual Evaluation. *arXiv preprint arXiv:1911.03064*.

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J. Paul. 2020. Multilingual Twitter Corpus and Baselines for Evaluating Demographic Bias in Hate Speech Recognition. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Marseille, France.

Christoph Hube, Maximilian Idahl, and Besnik Fetahu. 2020. Debiasing Word Embeddings from Sentiment Associations in Names. In *Proceedings of the International Conference on Web Search and Data Mining*, pages 259–267, Houston, TX.

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 Association for Computational Linguistics (ACL)*.

Matei Ionita, Yury Kashnitsky, Ken Krige, Vladimir Larin, Dennis Logvinenko, and Atanas Atanasov. 2019. Resolving gendered ambiguous pronouns with BERT. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, pages 113–119, Florence, Italy.

Hailey James-Sorenson and David Alvarez-Melis. 2019. Probabilistic Bias Mitigation in Word Embeddings. In *Proceedings of the Workshop on Human-Centric Machine Learning*, Vancouver, Canada.

Shengyu Jia, Tao Meng, Jieyu Zhao, and Kai-Wei Chang. 2020. Mitigating Gender Bias Amplification in Distribution by Posterior Regularization. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Taylor Jones, Jessica Rose Kalbfeld, Ryan Hancock, and Robin Clark. 2019. Testifying while black: An experimental study of court reporter accuracy in transcription of African American English. *Language*, 95(2).

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the Workshop on Noisy User-Generated Text*, pages 9–18, Beijing, China.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2016. Learning a POS tagger for AAVE-like language. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 1115–1120, San Diego, CA.

Pratik Joshi, Sebastian Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Jaap Jumelet, Willem Zuidema, and Dieuwke Hupkes. 2019. Analysing Neural Language Models: Contextual Decomposition Reveals Default Reasoning in Number and Gender Assignment. In *Proceedings of the Conference on Natural Language Learning*, Hong Kong, China.

Marie-Odile Junker. 2018. Participatory action research for Indigenous linguistics in the digital age. In Shannon T. Bischoff and Carmen Jany, editors, *Insights from Practices in Community-Based Research*, pages 164–175. De Gruyter Mouton.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating Dialectal Variability for Socially Equitable Language Identification. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 51–57, Vancouver, Canada.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1641–1650, Florence, Italy.

Saket Karve, Lyle Ungar, and João Sedoc. 2019. Conceptor debiasing of word representations evaluated on WEAT. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, pages 40–48, Florence, Italy.

Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Danielle Raz, and P.M. Krafft. 2020. Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 45–55, Barcelona, Spain.

Stephen Kemmis. 2006. Participatory action research and the public sphere. *Educational Action Research*, 14(4):459–476.

Os Keyes. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW).

Os Keyes, Josephine Hoy, and Margaret Drouhard. 2019. Human-Computer Insurrection: Notes on an Anarchist HCI. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, Glasgow, Scotland, UK.

Jae Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. 2020. Intersectional Bias in Hate Speech and Abusive Language Datasets. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Svetlana Kiritchenko and Saif M. Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, LA.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, 17(4):401–412.

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

Sonja L. Lanehart and Ayesha M. Malik. 2018. Black Is, Black Isn't: Perceptions of Language and Blackness. In Jeffrey Reaser, Eric Wilbanks, Karissa Wojcik, and Walt Wolfram, editors, *Language Variety in the New South*. University of North Carolina Press.

Brian N. Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the Workshop on Ethics in Natural Language Processing*, pages 30–40, Valencia, Spain.

Anne Lauscher and Goran Glavaš. 2019. Are We Consistently Biased? Multidimensional Analysis of Biases in Distributional Word Vectors. In *Proceedings of the Joint Conference on Lexical and Computational Semantics*, pages 85–91, Minneapolis, MN.

Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2019. A General Framework for Implicit and Explicit Debiasing of Distributional Word Vector Spaces. *arXiv preprint arXiv:1909.06092*.

Christopher A. Le Dantec, Erika Shehan Poole, and Susan P. Wyche. 2009. Values as Lived Experience: Evolving Value Sensitive Design in Support of Value Discovery. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, Boston, MA.

Nayeon Lee, Andrea Madotto, and Pascale Fung. 2019. Exploring Social Bias in Chatbots using Stereotype Knowledge. In *Proceedings of the Workshop on Widening NLP*, pages 177–180, Florence, Italy.

Wesley Y. Leonard. 2012. Reframing language reclamation programmes for everybody's empowerment. *Gender and Language*, 6(2):339–367.

Paul Pu Liang, Irene Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2019. Towards Debiasing Sentence Representations. In *Proceedings of the NeurIPS Workshop on Human-Centric Machine Learning*, Vancouver, Canada.

Rosina Lippi-Green. 2012. *English with an Accent: Language, Ideology, and Discrimination in the United States*. Routledge.

Bo Liu. 2019. Anonymized BERT: An Augmentation Approach to the Gendered Pronoun Resolution Challenge. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, pages 120–125, Florence, Italy.

Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2019. Does Gender Matter? Towards Fairness in Dialogue Systems. *arXiv preprint arXiv:1910.10486*.

Felipe Alfaro Lois, José A.R. Fonollosa, and Costa-jà. 2019. BERT Masked Language Modeling for Co-reference Resolution. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, pages 76–81, Florence, Italy.

Brandon C. Loudermilk. 2015. Implicit attitudes and the perception of sociolinguistic variation. In Alexei Prikhodkine and Dennis R. Preston, editors, *Responses to Language Varieties: Variability, processes and outcomes*, pages 137–156.

Anastassia Loukina, Nitin Madnani, and Klaus Zechner. 2019. The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–10, Florence, Italy.

Kaiji Lu, Peter Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018. Gender bias in neural natural language processing. *arXiv preprint arXiv:1807.11714*.

Anne Maass. 1999. Linguistic intergroup bias: Stereotype perpetuation through language. *Advances in Experimental Social Psychology*, 31:79–121.

Nitin Madnani, Anastassia Loukina, Alina von Davier, Jill Burstein, and Aoife Cahill. 2017. Building Better Open-Source Tools to Support Fairness in Automated Scoring. In *Proceedings of the Workshop on Ethics in Natural Language Processing*, pages 41–52, Valencia, Spain.

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W. Black. 2019. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 801–809, Minneapolis, MN.

Ramón Antonio Martínez and Alexander Feliciano Mejía. 2019. Looking closely and listening carefully: A sociocultural approach to understanding the complexity of Latina/o/x students' everyday language. *Theory Into Practice*.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 5270–5278, Hong Kong, China.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On Measuring Social Biases in Sentence Encoders. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 629–634, Minneapolis, MN.

Elijah Mayfield, Michael Madaio, Shrimai Prabhumoye, David Gerritsen, Brittany McLaughlin, Ezekiel Dixon-Roman, and Alan W. Black. 2019. Equity Beyond Bias in Language Technologies for Education. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, Florence, Italy.

Katherine McCurdy and Oğuz Serbetçi. 2017. Grammatical gender associations outweigh topical gender bias in crosslinguistic word embeddings. In *Proceedings of the Workshop for Women & Underrepresented Minorities in Natural Language Processing*, Vancouver, Canada.

Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, and Aram Galstyan. 2019. Man is to Person as Woman is to Location: Measuring Gender Bias in Named Entity Recognition. *arXiv preprint arXiv:1910.10872*.

Michela Menegatti and Monica Rubini. 2017. Gender bias and sexism in language. In *Oxford Research Encyclopedia of Communication*. Oxford University Press.

Inom Mirzaev, Anthony Schulte, Michael Conover, and Sam Shah. 2019. Considerations for the interpretation of bias measures of word embeddings. *arXiv preprint arXiv:1906.08379*.

Salikoko S. Mufwene, Guy Bailey, and John R. Rickford, editors. 1998. *African-American English: Structure, History, and Use*. Routledge.

Michael J. Muller. 2007. Participatory Design: The Third Space in HCI. In *The Human-Computer Interaction Handbook*, pages 1087–1108. CRC Press.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. 2013. "How Old Do You Think I Am?": A Study of Language and Age in Twitter. In *Proceedings of the Conference on Web and Social Media (ICWSM)*, pages 439–448, Boston, MA.

Malvina Nissim, Rik van Noord, and Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended Bias in Misogyny Detection. In *Proceedings of the Conference on Web Intelligence*, pages 149–155.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data*, 2.

Alexandra Olteanu, Kartik Talamadupula, and Kush R. Varshney. 2017. The Limits of Abstract Evaluation Metrics: The Case of Hate Speech Detection. In *Proceedings of the ACM Web Science Conference*, Troy, NY.

Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in word embeddings. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 446–457, Barcelona, Spain.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 2799–2804, Brussels, Belgium.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Xiangyu Peng, Siyan Li, Spencer Frazier, and Mark Riedl. 2020. Fine-Tuning a Transformer-Based Language Model to Avoid Generating Non-Normative Text. *arXiv preprint arXiv:2001.08764*.

Radomir Popović, Florian Lemmerich, and Markus Strohmaier. 2020. Joint Multiclass Debiasing of Word Embeddings. In *Proceedings of the International Symposium on Intelligent Systems*, Graz, Austria.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation Sensitivity Analysis to Detect Unintended Model Biases. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 5744–5749, Hong Kong, China.

Shrimai Prabhumoye, Elijah Mayfield, and Alan W. Black. 2019. Principled Frameworks for Evaluating Ethics in NLP Systems. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, Florence, Italy.

Marcelo Prates, Pedro Avelar, and Luis C. Lamb. 2019. Assessing gender bias in machine translation: A case study with google translate. *Neural Computing and Applications*.

Rasmus Précenth. 2019. Word embeddings and gender stereotypes in Swedish and English. Master's thesis, Uppsala University.

Dennis R. Preston. 2009. Are you really smart (or stupid, or cute, or ugly, or cool)? Or do you just talk that way? *Language attitudes, standardization and language change. Oslo: Novus forlag*, pages 105–129.

Flavien Prost, Nithum Thain, and Tolga Bolukbasi. 2019. Debiasing Embeddings for Reduced Gender Bias in Text Classification. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, pages 69–75, Florence, Italy.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically Neutralizing Subjective Bias in Text. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, New York, NY.

Arun K. Pujari, Ansh Mittal, Anshuman Padhi, Anshul Jain, Mukesh Jadon, and Vikas Kumar. 2019. Debiasing Gender biased Hindi Words with Word-embedding. In *Proceedings of the International Conference on Algorithms, Computing and Artificial Intelligence*, pages 450–456.

Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. 2019. Reducing gender bias in word-level

language models with a gender-equalizing loss function. In *Proceedings of the ACL Student Research Workshop*, pages 223–228, Florence, Italy.

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 Association for Computational Linguistics (ACL)*.

John R. Rickford and Sharese King. 2016. Language and linguistics on trial: Hearing Rachel Jeantel (and other vernacular speakers) in the courtroom and beyond. *Language*, 92(4):948–988.

Anthony Rios. 2020. FuzzE: Fuzzy Fairness Evaluation of Offensive Language Classifiers on African-American English. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, New York, NY.

Gerald Roche. 2019. Articulating language oppression: colonialism, coloniality and the erasure of Ti-betâĂŹs minority languages. *Patterns of Prejudice*.

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Tauman Kalai. 2019. What's in a Name? Reducing Bias in Bios without Access to Protected Attributes. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 4187–4195, Minneapolis, MN.

Jonathan Rosa. 2019. Contesting Representations of Migrant "Illegality" through the Drop the I-Word Campaign: Rethinking Language Change and Social Change. In Netta Avineri, Laura R. Graham, Eric J. Johnson, Robin Conley Riner, and Jonathan Rosa, editors, *Language and Social Justice in Practice*. Routledge.

Jonathan Rosa and Christa Burdick. 2017. Language Ideologies. In Ofelia García, Nelson Flores, and Massimiliano Spotti, editors, *The Oxford Handbook of Language and Society*. Oxford University Press.

Jonathan Rosa and Nelson Flores. 2017. Unsettling race and language: Toward a raciolinguistic perspective. *Language in Society*, 46:621–647.

Sara Rosenthal and Kathleen McKeown. 2011. Age Prediction in Blogs: A Study of Style, Content, and Online Behavior in Pre- and Post-Social Media Generations. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 763–772, Portland, OR.

Candace Ross, Boris Katz, and Andrei Barbu. 2020. Measuring Social Biases in Grounded Vision and Language Embeddings. *arXiv preprint arXiv:2002.08911*.

Richard Rothstein. 2017. *The Color of Law: A Forgotten History of How Our Government Segregated America*. Liveright Publishing.

David Rozado. 2020. Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types. *PLOS One*.

Elayne Ruane, Abeba Birhane, and Anthony Ventresque. 2019. Conversational AI: Social and Ethical Considerations. In *Proceedings of the Irish Conference on Artificial Intelligence and Cognitive Science*, Galway, Ireland.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 8–14, New Orleans, LA.

Elizabeth B.N. Sanders. 2002. From user-centered to participatory design approaches. In Jorge Frascara, editor, *Design and the Social Sciences: Making Connections*, pages 18–25. CRC Press.

Brenda Salenave Santana, Vinicius Woloszyn, and Leandro Krug Wives. 2018. Is there gender bias and stereotype in Portuguese word embeddings? In *Proceedings of the International Conference on the Computational Processing of Portuguese Student Research Workshop*, Canela, Brazil.

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 Association for Computational Linguistics (ACL)*, pages 1668–1678, Florence, Italy.

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 Association for Computational Linguistics (ACL)*.

Hanna Sassaman, Jennifer Lee, Jenessa Irvine, and Shankar Narayan. 2020. Creating Community-Based Tech Policy: Case Studies, Lessons Learned, and What Technologists and Communities Can Do Together. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain.

Danielle Saunders and Bill Byrne. 2020. Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Tyler Schnoebelen. 2017. Goal-Oriented Design for Ethical Machine Learning and NLP. In *Proceedings of the Workshop on Ethics in Natural Language Processing*, pages 88–93, Valencia, Spain.

Sabine Sczesny, Magda Formanowicz, and Franziska Moser. 2016. Can gender-fair language reduce gender stereotyping and discrimination? *Frontiers in Psychology*, 7.

João Sedoc and Lyle Ungar. 2019. The Role of Protected Class Word Lists in Bias Identification of Contextualized Word Representations. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, pages 55–61, Florence, Italy.

Procheta Sen and Debasis Ganguly. 2020. Towards Socially Responsible AI: Cognitive Bias-Aware Multi-Objective Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, New York, NY.

Deven Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Judy Hanwen Shen, Lauren Fratamico, Iyad Rahwan, and Alexander M. Rush. 2018. *Darling* or *Babygirl?* Investigating Stylistic Bias in Sentiment Analysis. In *Proceedings of the Workshop on Fairness, Accountability, and Transparency (FAT/ML)*, Stockholm, Sweden.

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 Empirical Methods in Natural Language Processing (EMNLP)*, pages 3398–3403, Hong Kong, China.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2020. Towards Controllable Biases in Language Generation. *arXiv preprint arXiv:2005.00268*.

Seungjae Shin, Kyungwoo Song, JoonHo Jang, Hyemi Kim, Weonyoung Joo, and Il-Chul Moon. 2020. Neutralizing Gender Bias in Word Embedding with Latent Disentanglement and Counterfactual Generation. *arXiv preprint arXiv:2004.03133*.

Jesper Simonsen and Toni Robertson, editors. 2013. *Routledge International Handbook of Participatory Design*. Routledge.

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

Yolande Strengers, Lizhe Qu, Qiongkai Xu, and Jarrod Knibbe. 2020. Adhering, Steering, and Queering: Treatment of Gender in Natural Language Generation. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, Honolulu, HI.

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 Association for Computational Linguistics (ACL)*, pages 1630–1640, Florence, Italy.

Adam Sutton, Thomas Lansdall-Welfare, and Nello Cristianini. 2018. Biased embeddings from wild data: Measuring, understanding and removing. In *Proceedings of the International Symposium on Intelligent Data Analysis*, pages 328–339, 's-Hertogenbosch, Netherlands.

Chris Sweeney and Maryam Najafian. 2019. A Transparent Framework for Evaluating Unintended Demographic Bias in Word Embeddings. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1662–1667, Florence, Italy.

Chris Sweeney and Maryam Najafian. 2020. Reducing sentiment polarity for demographic attributes in word embeddings using adversarial learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 359–368, Barcelona, Spain.

Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan, Mark D.M. Leiserson, and Adam Tauman Kalai. 2019. What are the biases in my word embedding? In *Proceedings of the Conference on Artificial Intelligence, Ethics, and Society (AIES)*, Honolulu, HI.

Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. 2020. It's Morphin' Time! Combating Linguistic Discrimination with Inflectional Perturbations. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Yi Chern Tan and L. Elisa Celis. 2019. Assessing Social and Intersectional Biases in Contextualized Word Representations. In *Proceedings of the Conference on Neural Information Processing Systems*, Vancouver, Canada.

J. Michael Terry, Randall Hendrick, Evangelos Evangelou, and Richard L. Smith. 2010. Variable dialect switching among African American children: Inferences about working memory. *Lingua*, 120(10):2463–2475.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A Report on the First Native Language Identification Shared Task. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pages 48–57, Atlanta, GA.

Mike Thelwall. 2018. Gender Bias in Sentiment Analysis. *Online Information Review*, 42(1):45–57.

Kristen Vaccaro, Karrie Karahalios, Deirdre K. Mulligan, Daniel Kluttz, and Tad Hirsch. 2019. Contestability in Algorithmic Systems. In *Conference Companion Publication of the 2019 on Computer*

Supported Cooperative Work and Social Computing, pages 523–527, Austin, TX.

Ameya Vaidya, Feng Mai, and Yue Ning. 2019. Empirical Analysis of Multi-Task Learning for Reducing Model Bias in Toxic Comment Detection. *arXiv preprint arXiv:1909.09758v2*.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting Gender Right in Neural Machine Translation. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 3003–3008, Brussels, Belgium.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Causal Mediation Analysis for Interpreting Neural NLP: The Case of Gender Bias. *arXiv preprint arXiv:2004.12265*.

Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez, and Caiming Xiong. 2020. Double-Hard Debias: Tailoring Word Embeddings for Gender Bias Mitigation. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Zili Wang. 2019. MSnet: A BERT-based Network for Gendered Pronoun Resolution. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, pages 89–95, Florence, Italy.

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 Workshop on Gender Bias in Natural Language Processing*, pages 1–7, Florence, Italy.

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*, 6:605–618.

Walt Wolfram and Natalie Schilling. 2015. *American English: Dialects and Variation*, 3 edition. Wiley Blackwell.

Austin P. Wright, Omar Shaikh, Haekyu Park, Will Epperson, Muhammed Ahmed, Stephane Pinel, Diyi Yang, and Duen Horng (Polo) Chau. 2020. RECAST: Interactive Auditing of Automatic Toxicity Detection Models. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, Honolulu, HI.

Yinchuan Xu and Junlin Yang. 2019. Look again at the syntax: Relational graph convolutional network for gendered ambiguous pronoun resolution. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, pages 96–101, Florence, Italy.

Kai-Chou Yang, Timothy Niven, Tzu-Hsuan Chou, and Hung-Yu Kao. 2019. Fill the GAP: Exploiting BERT for Pronoun Resolution. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, pages 102–106, Florence, Italy.

Zekun Yang and Juan Feng. 2020. A Causal Inference Method for Reducing Gender Bias in Word Embedding Relations. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, New York, NY.

Daisy Yoo, Anya Ernest, Sofia Serholt, Eva Eriksson, and Peter Dalsgaard. 2019. Service Design in HCI Research: The Extended Value Co-creation Model. In *Proceedings of the Halfway to the Future Symposium*, Nottingham, United Kingdom.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the Conference on Artificial Intelligence, Ethics, and Society (AIES)*, New Orleans, LA.

Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. 2020a. Demographics Should Not Be the Reason of Toxicity: Mitigating Discrimination in Text Classifications with Instance Weighting. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020b. Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning*.

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 Association for Computational Linguistics (ACL)*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender Bias in Contextualized Word Embeddings. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 629–634, Minneapolis, MN.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 2979–2989, Copenhagen, Denmark.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 15–20, New Orleans, LA.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning Gender-Neutral Word Embeddings. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 4847–4853, Brussels, Belgium.

Alina Zhiltsova, Simon Caton, and Catherine Mulwa. 2019. Mitigation of Unintended Biases against Non-Native English Texts in Sentiment Analysis. In *Proceedings of the Irish Conference on Artificial Intelligence and Cognitive Science*, Galway, Ireland.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical genders. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 5279–5287, Hong Kong, China.

Ran Zmigrod, S. J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1651–1661, Florence, Italy.

A Appendix

In Table 3, we provide examples of the papers' motivations and techniques across several NLP tasks.

A.1 Categorization details

In this section, we provide some additional details about our method—specifically, our categorization.

What counts as being covered by an NLP task?
We considered a paper to cover a given NLP task if it analyzed "bias" with respect to that task, but not if it only evaluated overall performance on that task. For example, a paper examining the impact of mitigating "bias" in word embeddings on "bias" in sentiment analysis would be counted as covering both NLP tasks. In contrast, a paper assessing whether performance on sentiment analysis degraded after mitigating "bias" in word embeddings would be counted only as focusing on embeddings.

What counts as a motivation? We considered a motivation to include any description of the problem that motivated the paper or proposed quantitative technique, including any normative reasoning.

We excluded from the "Vague/unstated" category of motivations the papers that participated in the Gendered Ambiguous Pronoun (GAP) Shared Task at the First ACL Workshop on Gender Bias in NLP. In an ideal world, shared task papers would engage with "bias" more critically, but given the nature of shared tasks it is understandable that they

do not. As a result, we excluded them from our counts for techniques as well. We cite the papers here; most propose techniques we would have categorized as "Questionable correlations," with a few as "Other representational harms" (Abzaliev, 2019; Attree, 2019; Bao and Qiao, 2019; Chada, 2019; Ionita et al., 2019; Liu, 2019; Lois et al., 2019; Wang, 2019; Xu and Yang, 2019; Yang et al., 2019).

We excluded Dabas et al. (2020) from our survey because we could not determine what this paper's user study on fairness was actually measuring.

Finally, we actually categorized the motivation for Liu et al. (2019) (i.e., the last row in Table 3) as "Questionable correlations" due to a sentence elsewhere in the paper; had the paragraph we quoted been presented without more detail, we would have categorized the motivation as "Vague/unstated."

A.2 Full categorization: Motivations

Allocational harms Hovy and Spruit (2016); Caliskan et al. (2017); Madnani et al. (2017); Dixon et al. (2018); Kiritchenko and Mohammad (2018); Shen et al. (2018); Zhao et al. (2018b); Bhaskaran and Bhallamudi (2019); Bordia and Bowman (2019); Brunet et al. (2019); Chaloner and Maldonado (2019); De-Arteaga et al. (2019); Dev and Phillips (2019); Font and Costa-jussà (2019); James-Sorenson and Alvarez-Melis (2019); Kurita et al. (2019); Mayfield et al. (2019); Pujari et al. (2019); Romanov et al. (2019); Ruane et al. (2019); Sedoc and Ungar (2019); Sun et al. (2019); Zmigrod et al. (2019); Hutchinson et al. (2020); Papakyriakopoulos et al. (2020); Ravfogel et al. (2020); Strengers et al. (2020); Sweeney and Najafian (2020); Tan et al. (2020); Zhang et al. (2020b).

Stereotyping Bolukbasi et al. (2016a,b); Caliskan et al. (2017); McCurdy and Serbetçi (2017); Rudinger et al. (2017); Zhao et al. (2017); Curry and Rieser (2018); Díaz et al. (2018); Santana et al. (2018); Sutton et al. (2018); Zhao et al. (2018a,b); Agarwal et al. (2019); Basta et al. (2019); Bhaskaran and Bhallamudi (2019); Bordia and Bowman (2019); Brunet et al. (2019); Cao and Daumé (2019); Chaloner and Maldonado (2019); Cho et al. (2019); Dev and Phillips (2019); Font and Costa-jussà (2019); Gonen and Goldberg (2019); James-Sorenson and Alvarez-Melis (2019); Kaneko and Bollegala (2019); Karve et al. (2019); Kurita et al. (2019); Lauscher and Glavaš (2019); Lee et al. (2019); Manzini et al. (2019); Mayfield

NLP task	Stated motivation	Categories	
		Motivations	**Techniques**
Language modeling (Bordia and Bowman, 2019)	*"Existing biases in data can be amplified by models and the resulting output consumed by the public **can influence them, encourage and reinforce harmful stereotypes**, or **distort the truth**. Automated systems that depend on these models can **take problematic actions** based on biased profiling of individuals."*	Allocational harms, stereotyping	Questionable correlations
Sentiment analysis (Kiritchenko and Mohammad, 2018)	*"Other biases can be inappropriate and result in **negative experiences** for some groups of people. Examples include, **loan eligibility and crime recidivism prediction systems**...and **resumé sorting systems** that believe that men are more qualified to be programmers than women (Bolukbasi et al., 2016). Similarly, sentiment and emotion analysis systems can also perpetuate and accentuate inappropriate human biases, e.g., systems that consider **utterances from one race or gender to be less positive** simply because of their race or gender, or customer support systems that **prioritize a call from an angry male over a call from the equally angry female**."*	Allocational harms, other representational harms (system performance differences w.r.t. text written *by* different social groups)	Questionable correlations (differences in sentiment intensity scores w.r.t. text *about* different social groups)
Machine translation (Cho et al., 2019)	*"[MT training] may incur an association of gender-specified pronouns (in the target) and gender-neutral ones (in the source) for lexicon pairs that frequently collocate in the corpora. We claim that this kind of phenomenon seriously threatens the fairness of a translation system, in the sense that it **lacks generality** and **inserts social bias** to the inference. Moreover, the input is **not fully correct** (considering gender-neutrality) and **might offend the users** who expect fairer representations."*	Questionable correlations, other representational harms	Questionable correlations
Machine translation (Stanovsky et al., 2019)	*"Learned models exhibit social bias when their training data **encode stereotypes not relevant for the task, but the correlations are picked up anyway**."*	Stereotyping, questionable correlations	Stereotyping, other representational harms (system performance differences), questionable correlations
Type-level embeddings (Zhao et al., 2018b)	*"However, embeddings trained on human-generated corpora have been demonstrated to inherit **strong gender stereotypes** that reflect social constructs....Such a bias substantially affects **downstream applications**....This concerns the practitioners who use the embedding model to build **gender-sensitive applications such as a resume filtering system or a job recommendation system** as the automated system may discriminate candidates based on their gender, as reflected by their name. Besides, biased embeddings may implicitly affect downstream applications used in our daily lives. For example, when searching for 'computer scientist' using a search engine...a search algorithm using an embedding model in the backbone tends to **rank male scientists higher than females'** [sic], **hindering women from being recognized** and further exacerbating the gender inequality in the community."*	Allocational harms, stereotyping, other representational harms	Stereotyping
Type-level and contextualized embeddings (May et al., 2019)	*"[P]rominent word embeddings such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) encode systematic biases against women and black people (Bolukbasi et al., 2016; Garg et al., 2018), implicating many NLP systems in scaling up social injustice."*	Vague	Stereotyping
Dialogue generation (Liu et al., 2019)	*"Since the goal of dialogue systems is to talk with users...if the systems show discriminatory behaviors in the interactions, the user experience will be adversely affected. Moreover, public commercial chatbots can get resisted for their improper speech."*	Vague/unstated	Stereotyping, other representational harms, questionable correlations

Table 3: Examples of the categories into which the papers' motivations and proposed quantitative techniques for measuring or mitigating "bias" fall. Bold text in the quotes denotes the content that yields our categorizations.

et al. (2019); Précenth (2019); Pujari et al. (2019); Ruane et al. (2019); Stanovsky et al. (2019); Sun et al. (2019); Tan and Celis (2019); Webster et al. (2019); Zmigrod et al. (2019); Gyamfi et al. (2020); Hube et al. (2020); Hutchinson et al. (2020); Kim et al. (2020); Nadeem et al. (2020); Papakyriakopoulos et al. (2020); Ravfogel et al. (2020); Rozado (2020); Sen and Ganguly (2020); Shin et al. (2020); Strengers et al. (2020).

Other representational harms Hovy and Søgaard (2015); Blodgett et al. (2016); Bolukbasi et al. (2016b); Hovy and Spruit (2016); Blodgett and O'Connor (2017); Larson (2017); Schnoebelen (2017); Blodgett et al. (2018); Curry and Rieser (2018); Díaz et al. (2018); Dixon et al. (2018); Kiritchenko and Mohammad (2018); Park et al. (2018); Shen et al. (2018); Thelwall (2018); Zhao et al. (2018b); Badjatiya et al. (2019); Bagdasaryan et al. (2019); Bamman et al. (2019); Cao and Daumé (2019); Chaloner and Maldonado (2019); Cho et al. (2019); Davidson et al. (2019); De-Arteaga et al. (2019); Fisher (2019); Font and Costa-jussà (2019); Garimella et al. (2019); Loukina et al. (2019); Mayfield et al. (2019); Mehrabi et al. (2019); Nozza et al. (2019); Prabhakaran et al. (2019); Romanov et al. (2019); Ruane et al. (2019); Sap et al. (2019); Sheng et al. (2019); Sun et al. (2019); Sweeney and Najafian (2019); Vaidya et al. (2019); Gaut et al. (2020); Gencoglu (2020); Hovy et al. (2020); Hutchinson et al. (2020); Kim et al. (2020); Peng et al. (2020); Rios (2020); Sap et al. (2020); Shah et al. (2020); Sheng et al. (2020); Tan et al. (2020); Zhang et al. (2020a,b).

Questionable correlations Jørgensen et al. (2015); Hovy and Spruit (2016); Madnani et al. (2017); Rudinger et al. (2017); Zhao et al. (2017); Burns et al. (2018); Dixon et al. (2018); Kiritchenko and Mohammad (2018); Lu et al. (2018); Park et al. (2018); Shen et al. (2018); Zhang et al. (2018); Badjatiya et al. (2019); Bhargava and Forsyth (2019); Cao and Daumé (2019); Cho et al. (2019); Davidson et al. (2019); Dev et al. (2019); Garimella et al. (2019); Garg et al. (2019); Huang et al. (2019); James-Sorenson and Alvarez-Melis (2019); Kaneko and Bollegala (2019); Liu et al. (2019); Karve et al. (2019); Nozza et al. (2019); Prabhakaran et al. (2019); Romanov et al. (2019); Sap et al. (2019); Sedoc and Ungar (2019); Stanovsky et al. (2019); Sweeney and Najafian (2019); Vaidya et al. (2019); Zhiltsova et al. (2019);

Chopra et al. (2020); Gonen and Webster (2020); Gyamfi et al. (2020); Hube et al. (2020); Ravfogel et al. (2020); Rios (2020); Ross et al. (2020); Saunders and Byrne (2020); Sen and Ganguly (2020); Shah et al. (2020); Sweeney and Najafian (2020); Yang and Feng (2020); Zhang et al. (2020a).

Vague/unstated Rudinger et al. (2018); Webster et al. (2018); Dinan et al. (2019); Florez (2019); Jumelet et al. (2019); Lauscher et al. (2019); Liang et al. (2019); Maudslay et al. (2019); May et al. (2019); Prates et al. (2019); Prost et al. (2019); Qian et al. (2019); Swinger et al. (2019); Zhao et al. (2019); Zhou et al. (2019); Ethayarajh (2020); Huang et al. (2020); Jia et al. (2020); Popović et al. (2020); Pryzant et al. (2020); Vig et al. (2020); Wang et al. (2020); Zhao et al. (2020).

Surveys, frameworks, and meta-analyses Hovy and Spruit (2016); Larson (2017); McCurdy and Serbetçi (2017); Schnoebelen (2017); Basta et al. (2019); Ethayarajh et al. (2019); Gonen and Goldberg (2019); Lauscher and Glavaš (2019); Loukina et al. (2019); Mayfield et al. (2019); Mirzaev et al. (2019); Prabhumoye et al. (2019); Ruane et al. (2019); Sedoc and Ungar (2019); Sun et al. (2019); Nissim et al. (2020); Rozado (2020); Shah et al. (2020); Strengers et al. (2020); Wright et al. (2020).

B Full categorization: Techniques

Allocational harms De-Arteaga et al. (2019); Prost et al. (2019); Romanov et al. (2019); Zhao et al. (2020).

Stereotyping Bolukbasi et al. (2016a,b); Caliskan et al. (2017); McCurdy and Serbetçi (2017); Díaz et al. (2018); Santana et al. (2018); Sutton et al. (2018); Zhang et al. (2018); Zhao et al. (2018a,b); Agarwal et al. (2019); Basta et al. (2019); Bhaskaran and Bhallamudi (2019); Brunet et al. (2019); Cao and Daumé (2019); Chaloner and Maldonado (2019); Dev and Phillips (2019); Ethayarajh et al. (2019); Gonen and Goldberg (2019); James-Sorenson and Alvarez-Melis (2019); Jumelet et al. (2019); Kaneko and Bollegala (2019); Karve et al. (2019); Kurita et al. (2019); Lauscher and Glavaš (2019); Lauscher et al. (2019); Lee et al. (2019); Liang et al. (2019); Liu et al. (2019); Manzini et al. (2019); Maudslay et al. (2019); May et al. (2019); Mirzaev et al. (2019); Prates et al. (2019); Précenth (2019); Prost et al. (2019); Pujari et al. (2019); Qian et al. (2019);

Sedoc and Ungar (2019); Stanovsky et al. (2019); Tan and Celis (2019); Zhao et al. (2019); Zhou et al. (2019); Chopra et al. (2020); Gyamfi et al. (2020); Nadeem et al. (2020); Nissim et al. (2020); Papakyriakopoulos et al. (2020); Popović et al. (2020); Ravfogel et al. (2020); Ross et al. (2020); Rozado (2020); Saunders and Byrne (2020); Shin et al. (2020); Vig et al. (2020); Wang et al. (2020); Yang and Feng (2020); Zhao et al. (2020).

Other representational harms   Jørgensen et al. (2015); Hovy and Søgaard (2015); Blodgett et al. (2016); Blodgett and O'Connor (2017); Blodgett et al. (2018); Curry and Rieser (2018); Dixon et al. (2018); Park et al. (2018); Thelwall (2018); Webster et al. (2018); Badjatiya et al. (2019); Bagdasaryan et al. (2019); Bamman et al. (2019); Bhargava and Forsyth (2019); Cao and Daumé (2019); Font and Costa-jussà (2019); Garg et al. (2019); Garimella et al. (2019); Liu et al. (2019); Loukina et al. (2019); Mehrabi et al. (2019); Nozza et al. (2019); Sap et al. (2019); Sheng et al. (2019); Stanovsky et al. (2019); Vaidya et al. (2019); Webster et al. (2019); Ethayarajh (2020); Gaut et al. (2020); Gencoglu (2020); Hovy et al. (2020); Huang et al. (2020); Kim et al. (2020); Peng et al. (2020); Ravfogel et al. (2020); Rios (2020); Sap et al. (2020); Saunders and Byrne (2020); Sheng et al. (2020); Sweeney and Najafian (2020); Tan et al. (2020); Zhang et al. (2020a,b).

Questionable correlations   Jurgens et al. (2017); Madnani et al. (2017); Rudinger et al. (2017); Zhao et al. (2017); Burns et al. (2018); Díaz et al. (2018); Kiritchenko and Mohammad (2018); Lu et al. (2018); Rudinger et al. (2018); Shen et al. (2018); Bordia and Bowman (2019); Cao and Daumé (2019); Cho et al. (2019); Davidson et al. (2019); Dev et al. (2019); Dinan et al. (2019); Fisher (2019); Florez (2019); Font and Costa-jussà (2019); Garg et al. (2019); Huang et al. (2019); Liu et al. (2019); Nozza et al. (2019); Prabhakaran et al. (2019); Qian et al. (2019); Sap et al. (2019); Stanovsky et al. (2019); Sweeney and Najafian (2019); Swinger et al. (2019); Zhiltsova et al. (2019); Zmigrod et al. (2019); Hube et al. (2020); Hutchinson et al. (2020); Jia et al. (2020); Papakyriakopoulos et al. (2020); Popović et al. (2020); Pryzant et al. (2020); Saunders and Byrne (2020); Sen and Ganguly (2020); Shah et al. (2020); Sweeney and Najafian (2020); Zhang et al. (2020b).

Vague/unstated   None.

Surveys, frameworks, and meta-analyses   Hovy and Spruit (2016); Larson (2017); McCurdy and Serbetçi (2017); Schnoebelen (2017); Basta et al. (2019); Ethayarajh et al. (2019); Gonen and Goldberg (2019); Lauscher and Glavaš (2019); Loukina et al. (2019); Mayfield et al. (2019); Mirzaev et al. (2019); Prabhumoye et al. (2019); Ruane et al. (2019); Sedoc and Ungar (2019); Sun et al. (2019); Nissim et al. (2020); Rozado (2020); Shah et al. (2020); Strengers et al. (2020); Wright et al. (2020).