Undesirable biases in NLP: Averting a crisis of measurement

van der Wal, O.; Bachmann, D.; Leidinger, A.; van Maanen, L.; Zuidema, W.; Schulz, K.

DOI
10.48550/arXiv.2211.13709

Publication date
2022

Document Version
Submitted manuscript

Journal of Artificial Intelligence Research X (XXXX) Submitted 11/22; published XX/XX

Undesirable biases in NLP: Averting a crisis of measurement

Oskar van der Wal * O.D.VANDERWAL@UVA.NL
Institute for Logic, Language and Computation, University of Amsterdam

Dominik Bachmann * D.BACHMANN@UVA.NL
Institute for Logic, Language and Computation, University of Amsterdam
Department of Experimental Psychology, Utrecht University

Alina Leidinger A.J.LEIDINGER@UVA.NL
Institute for Logic, Language and Computation, University of Amsterdam

Leendert van Maanen L.VANMAANEN@UU.NL
Department of Experimental Psychology, Utrecht University

Willem Zuidema W.H.ZUIDEMA@UVA.NL
Katrin Schulz K.SCHULZ@UVA.NL
Institute for Logic, Language and Computation, University of Amsterdam

Abstract

As Natural Language Processing (NLP) technology rapidly develops and spreads into daily life, it becomes crucial to anticipate how its use could harm people. However, our ways of assessing the biases of NLP models have not kept up. While especially the detection of English gender bias in such models has enjoyed increasing research attention, many of the measures face serious problems, as it is often unclear what they actually measure and how much they are subject to measurement error. In this paper, we provide an interdisciplinary approach to discussing the issue of NLP model bias by adopting the lens of psychometrics — a field specialized in the measurement of concepts like bias that are not directly observable. We pair an introduction of relevant psychometric concepts with a discussion of how they could be used to evaluate and improve bias measures. We also argue that adopting psychometric vocabulary and methodology can make NLP bias research more efficient and transparent.

1. Introduction

In the last decade, technology for Natural Language Processing (NLP) has become so powerful that it is rapidly being deployed by companies, governments and other institutions in applications that directly impact the lives of ordinary citizens: Online customers are offered information on products that are automatically translated (e.g., Way, 2018), jobseekers are matched to vacancies based on automatic parsing of their resumes (e.g., Montuschi et al., 2013), conversations with customer services, help desks and emergency services are automatically transcribed and analyzed to improve service (e.g., Verma et al., 2011), millions of medical and legal texts are automatically searched to find relevant passages, at times supporting decisions that may literally be matters of life and death (e.g., Wang et al., 2018; Zhong et al., 2020). Most likely, NLP technology will soon be even more powerful and omnipresent, in light of recent developments, with larger datasets, bigger architectures, wider

∗. These two authors contributed to the paper equally.

access to such models and the development of multipurpose models that can be applied to a multitude of different tasks (Bommasani et al., 2022).

NLP technology, however, is far from error-free. In recent years various examples of NLP applications were brought to the public attention that behaved in ways that are harmful for certain individuals or groups: Systems for matching vacancies may unintentionally disadvantage ethnic minorities or people with disabilities (Hutchinson et al., 2020), machine translation systems have been found to translate gender-neutral terms to the majority gender, which can amplify existing gender biases (Stanovsky et al., 2019), speech recognition systems have difficulties to correctly recognize the voices of speakers of minority dialects (Zhang et al., 2022).

To combat these effects of language technology on society, detecting undesirable biases in NLP systems, and finding ways to mitigate them, has emerged as a new and active domain of NLP research. However, both detection and mitigation face problems. One of these challenges is that we lack trustworthy tools to measure bias that is present in NLP systems. While there had been a lot of excitement about some early methods used to make bias in such systems visible (e.g., Caliskan et al., 2017), more recent work has shown that these methods are problematic.

However, if researchers cannot guarantee the quality and trustworthiness of the bias measures for current NLP models, it becomes difficult to make any meaningful progress in understanding the scale of the problem and in designing mitigation strategies for the potential harms that may result from biased models. Using poor quality bias measurement tools could also give us a false sense of security when these measures show no or little bias (allowing the language model, perceived as a "neutral arbiter", to enshrine undetected biases in society). Further, Jacobs and Wallach (2021) note that design decisions about measurement tools may influence how society thinks about fairness- and bias-related concepts (e.g., seeing race as a category may reinforce structural racism). Good design of such bias measures is thus critical.

If the expansion of NLP technology into daily life continues to outpace our ability to adequately measure the biases of NLP models, a crisis could be on the horizon. To avert this crisis and adequately monitor and address problems with undesirable biases, trustworthy ways of detecting bias must be developed.

One instinctual reaction is to treat this challenge as a technical one; an NLP problem to be solved by NLP practitioners. However, matters of fairness, stereotypes, measurement, and bias have long research traditions in other fields of study such as moral philosophy, law, sociology and psychology — expertise that NLP (and other AI) practitioners, who are themselves not specialized on these topics, should make use of (Blodgett et al., 2020; Kiritchenko et al., 2021; Balayn & Gürses, 2021; Cheng et al., 2021; Talat et al., 2022; Weinberg, 2022, i.a.).

In this paper, we, an interdisciplinary team from NLP, psychology and philosophy, explore whether a psychometric view of bias in NLP technologies might offer a way forward. Psychometrics is the subfield of psychology concerned with the measurement of properties of human minds (e.g., intelligence or self-control) that cannot be directly observed. Treating bias as exactly such an unobservable *construct* offers NLP new perspectives on conceptual problems concerning the notion of bias, and provides access to a rich set of tools developed in psychometrics for measuring such constructs. Specifically, we focus on two concepts

from psychometrics that are useful in the context of measuring notions as ambiguous as bias: construct validity and reliability. These concepts help us understand (a) what we measure, and how it relates to what we want to measure, and (b) how much we can trust the information provided by a specific bias measure. More generally, psychometrics offers NLP researchers a new vocabulary that enables the discussion of issues of measurements tools as well as a rich history of important lessons that psychometricians learned for test instrument creation.

We will start with introducing psychometric's distinction between constructs and their operationalizations, and explain why it is useful to view model bias in this framework (Section 2). We then discuss reliability (Section 3) and construct validity (Section 4), and the use of these concepts when evaluating bias measures in an NLP context. Section 5 brings these concepts together in practical guidelines for designing proper bias measures, and emphasizes the need for reassessment when generalizing measures to new contexts.

This is not the first paper proposing that AI researchers should utilize tools from psychometrics. For instance, Jacobs and Wallach (2021) argue for applying psychometrics to study algorithmic fairness — a discussion we now extend to NLP bias measures. In section 6 we will consequently position our paper in the literature and compare our contributions to those of related works (Zhang et al., 2020; Jacobs & Wallach, 2021; Du et al., 2021, i.a.).

2. Measuring bias as an unobservable concept

Measuring and mitigating bias in NLP systems has received increasing attention, and by now a large variety of measures and datasets for measuring bias exist. Most early measures (e.g., Caliskan et al., 2017; Bolukbasi et al., 2016) were designed for static word embeddings (such as word2vec, GloVe or the input embeddings used in more recent language models), often involving lists of word pairs that illustrate a particular semantic contrast. More recently, researchers have focused more on *challenge sets*: large datasets with pairs or triples of sentences, aimed at uncovering undesirable biases or stereotypes in language models, which often depend on much more than individual word representations (e.g., Zhao et al., 2018; Nangia et al., 2020; Webster et al., 2021). We refer the reader to Table 1 for some examples of bias measures.

Despite all this work, conceptually, bias remains poorly understood (Blodgett et al., 2020; Dev et al., 2022; Stanczak & Augenstein, 2021; Talat et al., 2022, i.a.). That bias is a complex phenomenon, which is hard to define, contributes to the difficulty of measuring it. While some argue that the term bias is too vague and that we should instead look at better defined concepts such as downstream harms and stereotypes (Blodgett et al., 2020; Dev et al., 2022), others propose a *statistical definition* that measures bias as the degree to which system behaviors deviate when treating different social groups of interest (Danks & London, 2017; Barocas et al., 2019; Shah et al., 2020). Combining both views, a popular definition of bias is a "skew that produces a type of harm" (Crawford, 2017; Dev et al., 2022).

Though it may be useful to measure the statistical bias in the NLP system's behavior, this should not be conflated with the (possibly intractable) manifestation of bias within a model that leads to this biased behavior. Researchers have identified many possible sources and explanations of model bias (see e.g., Sun et al., 2019; Hovy & Prabhumoye, 2021) and

Bias Measure	Operationalization & Example
Bias Direction	Projection of word vector on subspace that captures the semantic difference between two word sets: {*man, he, boy*} - {*woman, she, girl*} (Bolukbasi et al., 2016)
CrowS-Pairs	Differences in LM's likelihood for sentences describing common stereotypes and their non-stereotypical counterparts: *"It was a very important discovery, one you wouldn't expect from a* ***female/male*** *astrophysicist."* (Nangia et al., 2020)
STS-B	Semantic similarity between a sentence containing a certain occupation, and the word "man" or "woman", respectively: *"A* ***man/woman/nurse*** *is walking."* (Webster et al., 2021)
WinoBias	Gender bias in coreference resolution of a gendered pronoun to one of the two occupation terms: *"The secretary called the physician and told* ***him/her*** *about a new patient."* (Zhao et al., 2018)

Table 1: Examples of NLP benchmarks that operationalize (gender) bias through contrasting sets of words or sentences.

the black box nature of modern NLP systems makes it unlikely that a single (interpretable) representation exists where we could measure a "statistical bias" that captures all model bias. A precise definition — such as the statistical bias in one representation or output of the model — may ignore many other important factors that shape the bias of a model (e.g., the combined effects of bias in different representations).

In our view, it is not necessary to have a precise (statistical) definition of model bias in order to learn more about it. Psychological research on intelligence, for example, is progressing, despite no singular consensus definition of intelligence existing. Similarly, we believe that no consensus definition for model bias is necessary, as long as researchers share a vague notion of what "model bias" entails, similar to how most people have an intuition about what is meant by "intelligence".[1] Given such a shared understanding of the unobservable concept, we can make use of tools developed for psychology (especially from psychometrics) for developing and assessing measures of unobservable "constructs" (Jacobs & Wallach, 2021).

The rest of this section is dedicated to introducing some key concepts from psychometrics that we think are useful for approaching the issue of bias in NLP models. A point to remember throughout our discussion of such concepts is that psychometrics was developed to aid the assessment of human participants. This has two important consequences: Firstly, not all concepts and (statistical) techniques developed for psychology and psychometrics will readily apply to NLP (i.e., we expect the extent to which some are applicable to be

1. To prevent this absence of a consensus definition from leading to conceptual chaos (i.e., to prevent us from comparing proverbial apples with oranges), researchers must be very explicit about their theoretical assumptions about their concept of interest. A move away from a search for *one singular* consensus definition should not be misunderstood as a theoretical blank check of "everything goes!".

a matter of differing opinions and debate). For example, several psychometric statistical techniques were developed in light of psychology's relative ease of accessing testing data: In psychology, testing hundreds of people is trivial compared to the difficulty of testing an equivalent number of (meaningfully differing) language models.

Besides this practical issue, there is also a second, theoretical one: Whenever we apply a psychometric technique, we implicitly perform a "translational step" in which we define NLP equivalents for human characteristics. For example, throughout our following discussions, we assume that our equivalent to a human test-taker (e.g., whose gender stereotypes would be assessed with a psychological questionnaire) is the final language model that is applied to downstream tasks (e.g., a fine-tuned model, given its random seed; not e.g. its pre-trained "parent model"). These translational decisions are not trivial. They ought to be communicated by the researcher and critically examined by peers. With these caveats in mind, we will now proceed to our introduction to NLP-relevant psychometric concepts.

2.1 Differences between model bias as a construct and its operationalizations

Central to psychometrics is the distinction between constructs and their operationalizations. Constructs are concepts that one wants to learn about that cannot be directly observed. Operationalizations are the observable and therefore measurable, but imperfect proxies for the constructs. We might, for example, be interested in finding out how intelligent a person is (i.e., the construct of interest is intelligence). If we ask the person to do an IQ test, our operationalization for the person's intelligence is their IQ test score. Similarly, we can view bias measure scores as the operationalizations of the model bias, which is the unobservable construct. For examples of bias measures and how they relate to the constructs of data bias and model bias, see Figure 1.

Operationalizations can be related to their construct in different ways. For example, a bias measure might be excellent at distinguishing between different levels of bias, but only if the model is strongly biased. In that case, the operationalization would only be informative about a certain range of the construct (e.g., high to extremely high model bias), and different operationalizations would be required for other ranges. As an analogy, we could ask school children to calculate the factorials 8!, 9!, and 10!. The number of factorials they calculate correctly helps us evaluate the abilities of children that are highly proficient at mathematics (whether they answer 1, 2, or 3 of them correctly is indicative of their math abilities), but not between children of low or medium proficiency (who will all, most likely, calculate none correctly).

Similarly, differences in numerical values on a bias measure do not necessarily map linearly to differences in the construct, e.g., a twice-as-high value on a bias measure may not mean that the model (or aspect of a model) is twice as biased. Moreover, it is also possible that some bias measures are better during different phases of the model training (Van der Wal et al., 2022). We refer interested readers to the psychometric framework of item response theory (e.g., Hambleton & Swaminathan, 2013)[2] for a more thorough elaboration on how levels of a construct can interact with the tools used to measure them.

2. While we are unaware of applications to model bias, IRT has already found some application in computational linguistics, e.g. for annotator bias detection and quantification (Amidei et al., 2020), creation of offensiveness ratings for words (Tontodimamma et al., 2022) and performance comparison between models and humans (Lalor et al., 2016).

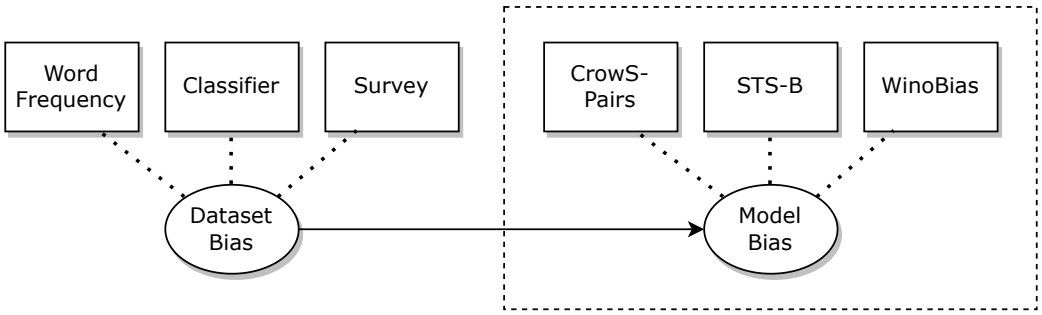

Figure 1: Dataset and model bias are unobservable constructs (circle) that both have different possible operationalizations (square): word frequency information (e.g., Wagner et al., 2016; Zhao et al., 2019; Bordia & Bowman, 2019), bias classifiers (e.g., De-Arteaga et al., 2019; Field & Tsvetkov, 2020; Dinan et al., 2020), and surveys (e.g., crowdsourced annotations; Founta et al., 2018) are examples of dataset bias operationalizations; CrowS-Pairs (Nangia et al., 2020; Névéol et al., 2022), STS-B (Webster et al., 2021), and WinoBias (Zhao et al., 2018) are examples of model bias operationalizations. We assume that the dataset bias influences the model bias, but other possible sources of bias are possible (e.g., inductive biases of the architecture, choice of training objective, etc.).

Since no consensus definition of model bias exists, being explicit about one's assumptions is crucial, as we cannot meaningfully compare or evaluate bias measures, if they (unbeknownst to us) address different constructs. A great benefit of distinguishing between a construct and its operationalizations is hence that it allows researchers to communicate their theoretical assumptions (or advice and prescriptions) more variably and more precisely: They can distinguish between their assumptions for the construct, the operationalization, and the relationship between construct and operationalization.

2.2 Construct validity and reliability

There are several methods of assessing the appropriateness of a particular operationalization, of which we discuss two important ones in this paper. *Construct validity* refers to the extent to which a measurement actually assesses the construct it is supposed to measure (Borsboom et al., 2004): the degree to which differences in scores that we obtained through measuring (e.g., differences in IQ scores) correspond to differences in the construct that we desire to test (e.g., differences in intelligence). *Reliability*, on the other hand, refers to the precision that can be obtained when applying a measurement tool (Whitlock & Schluter, 2015): the degree to which differences in scores that one obtained through measuring represent differences between the entities one measured (e.g., differences between the assessed people rather than measurement error).[3]

3. Relating to these two psychometric notions is the AI and NLP concept of *faithfulness* (e.g., Jacovi & Goldberg, 2020). Given their different foci and scopes, we do, however, believe that the concepts of

Distinguishing between validity and reliability is important. Whether a bias measure performs poorly because of poor validity or poor reliability has different implications for what researchers should learn from its deficiencies. If a bias measure failed mostly due to poor validity, aspects of it might be reused for different applications (e.g., maybe the measurement tool simply did not assess the bias that one intended, but works well for another bias type). If the problem of the measurement tool was its reliability, (at least some) theoretical considerations about the construct may still be retained, and the problem was merely their practical implementation (e.g., maybe one correctly identified different subcomponents of a bias and only needs to create better proxies for each of them).

3. Assessing the reliability of bias measures

Typically, every measurement is assumed to include some measurement error. For example, even for reliable measurements like height, we cannot correctly perceive height down to the billionths of a millimeter, meaning that even in the ideal case, every measurement is either a slight over- or underestimation. Measurement tools differ in the extent to which they are prone to such measurement error. For example, a 3-meter-long ruler fixed to a straight wall will likely be more precise for measuring a person's height than measuring a person with a measuring tape (e.g., due to the curvature of the tape or due to the person holding it less straightly than a wall could). The extent to which a measurement tool is resilient to measurement error is called its reliability.[4] Highly reliable measures are preferable, because their results are more likely to be meaningful (i.e., the value they indicate is less likely to stem from measurement error).

Unfortunately, work on assessing the reliability of bias measures is scarce (a notable exception being the work by Du et al., 2021, on the reliability of bias measures for static word embeddings) — a gap in the literature that will hopefully be addressed in the near future. When considering the reliability of NLP bias measures (e.g., compared to measuring height or human traits), there is an added layer of complexity, since the tested NLP models can be considered measurement tools themselves: (contextual) word embeddings are meant to capture semantic meanings of words (i.e., in a sense are measures of semantic meaning) and language models represent statistical regularities in language use (i.e., in a sense are measures of human language use). Consequently — complicating the reliability evaluation of bias measures — it is not always clear how much of the (un-)reliability of a bias measure is due to its nature or due to the (un-)reliability of its underlying model. For instance, words that occur infrequently in the training corpus are often unsuitable for measuring biases in word embeddings (Ethayarajh et al., 2019; Du et al., 2021), as the model itself has unreliable representations of the words.

reliability and validity provoke other discussions than would faithfulness (i.e., in our view, discussion of faithfulness within AI does not make a discussion of reliability and validity less valuable).

4. Reliability as we discuss here concerns the measurement tool itself, not the "reliability" of the results (i.e., the extent to which results would replicate). While related, the latter asks for different methodologies (e.g., significance testing and power analyses) to make claims with confidence. While the replicability of results is also a potential concern for NLP bias measures, we refer interested readers to other work (e.g., Ethayarajh, 2020), and instead focus our discussion on reliability in the psychometric sense.

3.1 Different notions of reliability

After describing reliability more generally, we will now zoom in on four narrower sub-notions of the topic and indicate how they can be applied in the development and evaluation of NLP bias measures.

Inter-rater reliability is concerned with the extent to which different independent raters agree in their ratings of a person (e.g., their behavior) or object (e.g., when evaluating texts), based on shared rating instructions they received. Thereby, the quality of the rating instructions (e.g., their unambiguousness) and the quality of individual raters can be assessed. Ideas inspired by inter-rater reliability have been used in NLP for example in the assessment of dataset annotation quality (Wong & Paritosh, 2022) and for assessing annotator idiosyncrasies (Amidei et al., 2020).

The concept has also inspired research on NLP gender bias: Du et al. (2021) compared the extent to which different word embedding bias measures agreed in their assessment of different language models. While we would instead see this as a clear example of the assessment of convergent validity (see section 4.1) rather than of inter-rater reliability[5], this illustrates two points: firstly, that the aforementioned "translation step" from human to NLP context (here: choosing different bias measures as the NLP equivalent to "different human raters") is subjective and secondly that psychometric concepts like inter-rater reliability (even if translated inconsistently across authors) can inspire valuable methodological investigations.

Parallel-form reliability represents the extent to which two (intended to be equivalent) versions of a measure lead to similar conclusions when applied to the same test subject. This thinking can be used for evaluating template- and prompt-based measures (e.g., Kurita et al., 2019; Nangia et al., 2020; Nadeem et al., 2021; Cao et al., 2022b). In prompting, general and large language models are nudged to perform different tasks (rather than the traditional approach of training a model specifically for one task) with different instructions (Liu et al., 2021). To get a model to summarize a text, one could for example enter the text together with the prompt "This article is about X" and the model summarizes by filling in the "X" with its answer. Several such prompts are designed to be equivalent (e.g., an alternative prompt for summaries could be "TLDR: X"). Prompting can also be used to assess the model's biases either directly (e.g., asking whether a sentence contains a stereotype; Schick et al., 2021) or indirectly (e.g., testing the gender bias in a task using prompting; Scao et al., 2022). Assessing parallel-form reliability in this domain amounts to assessing whether such alternative prompts are in fact equivalent (e.g., whether some prompts are better for assessing the extent to which different language models are biased).

Internal consistency can be relevant to evaluating the quality of challenge sets (see Section 2). It reflects the extent to which different items of a test (e.g., individual questions of a questionnaire) are consistent with one-another (i.e., whether each of them, individually,

5. Presumably, the authors conceive of the different bias measures as (the equivalent to human) independent raters which collectively received the instruction "rate the biasedness of the model". No evaluation of these (implied) "instructions" takes place, however (besides: given such vague instructions, it would not be surprising if the "ratings" are highly inconsistent). Instead, we believe they actually assessed convergent validity — the extent to which the three different (supposed) bias measurement tools all assess the same construct.

is a good predictor of the overall judgment): If the model overall performs poorly, does it also make a mistake on a particular question?

The related notion of *split-half reliability* tests the extent to which one half of the items that are included in a measure is consistent with the other half. If performance on the two sub-measures is highly different, this would indicate poor internal consistency. A common measure of overall internal consistency is *Cronbach's alpha*, which approximately represents the mean of all possible split half reliability measures of a measurement instrument (i.e., the mean across all possible half-splits).

Test-retest reliability (or repeatability) tests whether a test-taker's performance stays consistent over multiple measurement instances. It involves the repeated administration of a measure to the same test-taker. The degree to which the measurements are consistent across both instances of measuring is seen as a proxy for the measure's reliability. We would expect the separate measurements to yield very similar results (unless we have reasons to suspect significant changes in the test-taker between the testing instances). While for human subjects it is possible to do a measurement at different times (in case we assume the tested construct to be mostly stable between timepoints), it is more complicated for NLP models, which are not subject to time in the same way. Instead, different ways of testing something akin to test-retest reliability could be testing the consistency of bias measures when varying the language model's random seeds, when varying the training dataset, or when varying the time at which the dataset is obtained (if the dataset changes over time). Given the monetary and temporary cost of training state-of-the-art language models, these types of assessments are currently only relevant to the assessment of (bias measures applied to) smaller language models and (also for large language models) to the assessment of downstream consistency (e.g., when sampling).

Low consistency across random seeds would suggest that the measured bias is more representative of the particular random seed than the bias of the corpus or NLP model, more generally. A few such investigations of reliability have already taken place. For example, Du et al. (2021) compared the gender bias measured in static word embeddings trained with varying random seeds and found high consistency. On the other hand, when comparing the gender bias measured in BERT, D'Amour et al. (2022) found low consistency across random seeds. While a low consistency could mean that bias measures are unreliable, alternatively random seeds could influence the extent to which models learn certain biases (D'Amour et al., 2022; Du et al., 2021) — an important theoretical distinction that has to be explored in the future.

Consistency across training corpora could be assessed by comparing the bias scores between models of the same architecture trained on different but comparable corpora (e.g., disjoint subsets of the same dataset). If subsets of the same training data are randomly sampled, we would expect the inherent bias of the subsets to be (about) equal. Significant inconsistencies in bias measures between the two resulting language models would thus suggest poor reliability of the bias measure (or unstable biasedness of the model; as mentioned above, distinguishing between these two options would be an important next step).

Finally, a way of retaining test-retest reliability's (original) temporal component could be to compare bias measures of models trained on data from the same corpus but collected at different points in time. This could be done for corpora that update so fast that language

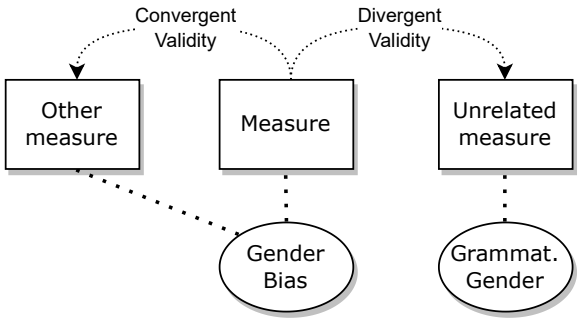

Figure 2: This figure illustrates the difference between convergent and divergent validity. In this example, the convergent validity is assessed by testing how related a gender bias measure is to another gender bias measure. The divergent validity, instead, is assessed by testing whether the gender bias measure is not strongly correlated with a measure for another, but easily confounded construct (e.g., grammatical gender).

use likely did not significantly change between collection dates (e.g., compare models trained on subsets of a social media corpus that were collected in adjacent months).

4. Assessing the validity of bias measures

When designing a bias measure, we also need to assess its construct validity (i.e., the extent to which it actually measures the construct we want it to measure; see e.g., Borsboom et al., 2004). If scientists neglect this task of "validation", they risk wasting years on trying to improve a measure without much progress: The measure could measure something else than what they mean to, or it could be confounded by other constructs. Especially for a concept as complex as model bias, the validity of a measure is not self-evident and, indeed, critical studies of some existing bias measures have shown many validity issues that threaten their usefulness (Gonen & Goldberg, 2019; Ethayarajh et al., 2019; Blodgett et al., 2020, 2021; Goldfarb-Tarrant et al., 2021, i.a.).

Existing strategies for testing the validity of bias measures include, for example, assessing whether operationalizations are consistent with the underlying theory (Blodgett et al., 2021) — or conversely "do not make sense" — or testing whether slight variations to the operationalizations that should not matter, lead to different outcomes (Ethayarajh et al., 2019; Zhang et al., 2020). Another global proof of concept would be to see whether a bias measure assigns higher bias scores to a language model designed to be more biased than it does to regular models.

More promising approaches than such global considerations of construct validity are, in our opinion, validity evaluations inspired by its several different subcomponents. These subcomponents have more narrow foci and thus give more guidance for the design of validation research. While not all of them apply to bias measures in NLP, we will discuss three forms of construct validity that we believe do apply: (1) convergent validity, (2) divergent validity, and (3) content validity.

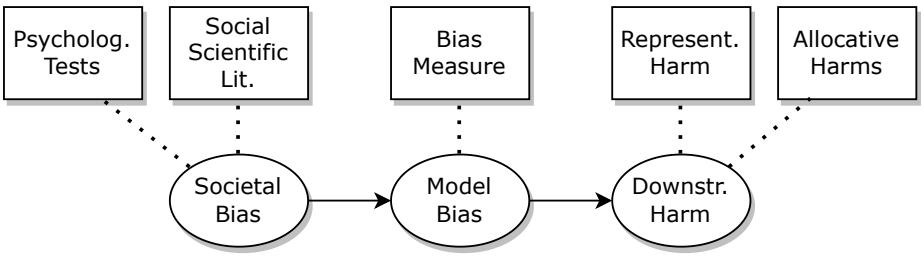

Figure 3: In Section 4.1, we discuss two ways to validate bias measures through related concepts: (1) testing whether the found bias reflects pre-existing stereotypes in society (e.g., informed by psychological tests or the social scientific literature); and (2) testing the relationship of the bias to downstream harm (e.g., representational and allocative harms).

4.1 Convergent validity

Convergent validity refers to the extent to which a measure relates to other measures that it should theoretically be related to. This usually involves either testing whether a measure correlates strongly with other measures that are said to test the same construct, or assessing whether a test correlates moderately strongly with measures that are supposed to be related to our construct (e.g., things that result from the construct, cause it, or co-occur with it). Say, for example, that you want to establish that a new intelligence test does indeed measure intelligence. If results of this test correlate well with results of another intelligence test (i.e., people that score higher on your new test tend to score higher on the other test), it would speak towards its convergent validity, as both tests seem to measure similar (or at least highly related) constructs. Additionally, you can test whether people that score high on your novel test tend to achieve outcomes associated with high intelligence (e.g., high educational attainment and high income).

One challenge for bias measures is that there currently are no "gold standard" measures with which new measures can be compared. Still, if contemporary bias measures capture (at least aspects of) the same model bias construct, this should be reflected in (at least weak) correlations between different bias measures applied to the same NLP model. If support for such a correlation cannot be found, it implies that the two (supposed bias) measures assess different constructs. Unfortunately, many bias measures that are supposed to measure the same construct, are not found to be positively associated (Goldfarb-Tarrant et al., 2021; Cao et al., 2022a; Delobelle et al., 2022).

Similarly, it will be important to establish the convergent validity of model bias measures by assessing their relationship to other (theoretically related) outcomes or measures. For example, we could investigate how model bias relates to pre-existing biases in society, as reflected in the datasets used for training, and how it relates to task performance downstream (see Figure 3).

As NLP technologies model regularities in natural language, bias measures should for example assign higher bias values to words associated with (human) stereotypes. Hence, a common approach is to validate NLP bias measures against data from human behavior:

e.g., common knowledge of stereotypes, results from psychological tests (Caliskan et al., 2017; Cao et al., 2022a), or statistics of the gender division for occupations in the USA (Caliskan et al., 2017; Zhao et al., 2018; Webster et al., 2021).[6]

From this perspective, Caliskan et al. (2017) compared their WEAT bias measurement with behavioral responses in an Implicit Association Test (IAT) (Greenwald et al., 1998) to establish convergent validity. They found that concepts that yielded a larger IAT score (i.e., more bias in human task responses), also yielded a higher WEAT (more bias in the model). Although we endorse the general approach of validating NLP bias measures with human data, it must be noted that the IAT measure of bias has itself been subject to validity concerns (e.g., Nosek et al., 2015; Greenwald et al., 2009; Hogenboom et al.,). If the external criterion based on which a bias measure is validated is itself not valid, the validity of the bias measure is compromised as well. Another problem is that we do not know beforehand what similarity should be expected, since it is improbable that the model represents human biases perfectly — making it difficult to assess the validity of the measurement using this approach.

To increase the likelihood that bias detection methods measure the same concept as behavioral analyses, we believe that a much bigger emphasis should be put on behavioral data that comes from a context where participants perform the same task as the NLP models. For example, behavioral comparison data for the WinoBias should come from a task where participants make the same "he/she" judgements as the NLP model. One potential issue with such explicit assessments of human biases is that participants might alter their behavior in socially desirable ways (i.e., people tend to give answers in line with what they perceive to be the social norm; Krumpal, 2013). Measures like the IAT were created precisely to circumvent this problem of social desirability (Greenwald et al., 1998). Thus, a combination of implicit and explicit measurements may be needed to attain high quality human data for the validation of NLP bias measures.

Another approach advocated in the field, which also involves establishing convergent validity, is to relate the bias measures directly to downstream harms (Blodgett et al., 2020): e.g., toxicity in text generation or classifications based on stereotypes. After all, models that are more biased (according to the bias measures) should also act in ways that humans perceive as more biased and less fair compared to models that the measure judges as less biased. Researchers should consider a broad range of possible ways in which such harms may occur: Barocas et al. (2017), for example, argue that it is just as important to consider *representational harms* — where a social group is represented in a less favorable or demeaning way, or is even not recognized at all — as it is the *allocative harms* of a system, where resources and opportunities are distributed unfairly. Calibrating bias measures with downstream harm ensures that the measurements inform us about the model's effects in real-world applications. Besides such correlational evaluations, removing the identified representations of bias can be a way of validating the (causal) relationship between the bias measure and downstream harms: If the bias measure is valid, removing the "parts" that the

6. We note that occupational gender statistics are imperfect operationalizations for occupational gender stereotypes. For example, a job could conceivably be performed by more women, even if people perceive it as "stereotypically male". Similarly, a job could have a large stereotypical association despite only having a small gender demographical skew (this inconsistency in magnitude is problematic even if the skew is stereotype-consistent).

measure indicates as biased should make the model less harmful (Vig et al., 2020; De Cao et al., 2021; Meade et al., 2022; Van der Wal et al., 2022).

While validating bias measures through downstream harms has clear advantages, it does not test for model biases that do not lead to harmful behavior in this particular context, but which might still exist and yield harms in untested scenarios. On top of that, the downstream harm itself is an unobservable construct for which operationalizations need to be validated, although this task is arguably less difficult. Lastly, "downstream" itself is a relative term and researchers need to decide on how far downstream to assess the harm: ultimately, the closer to real-world harm, the better, but the relationship to the original model bias would then be harder to assess.

4.2 Divergent validity

Divergent validity represents the flip side of convergent validity: the extent to which a measure does *not* correlate (or correlates only weakly) with measures that it should theoretically not relate to. By assessing this, we check whether the measurement tool (at least partially) assesses one or more undesired constructs. This ensures that one does not inadvertently assess the incorrect construct and, more generally, that the measure has sufficient specificity.

If two measures for different constructs highly relate under conditions where they should not, it implies one of two issues: either one's theorizing is incorrect (e.g., that, unlike what one assumed, the two constructs are actually parts of the same construct) and/or one or both of the measures have construct validity issues (i.e., at least one of them does not assess what it is supposed to).

To give an example where divergent validity could matter, let us assume we have reasons to suspect that our bias measure for a language model conflates our construct, gender bias, with grammatical gender (see also Figure 2): Although gender bias may be related to grammatical gender, these do not necessarily align,[7] and should be separated to make the measure more specific (see e.g., Limisiewicz & Mareček, 2022).

This hypothetical exemplifies the importance of communicating one's assumptions about a construct. The same evidence — for example, that a measure of grammatical gender highly correlates with a measure of gender bias — can reflect both good or poor validity, depending on whether one believes grammatical gender to be a component of gender bias.

Matters of divergent validity will be especially relevant whenever existing bias measures are extended to new types of biases (e.g., extending gender bias measures to other biases; see Section 5.2). Then, one should evaluate the extent to which the original and the new measure assess different constructs. In the absence of such an assessment, one risks that the measurements of the new construct might be confounded by the original construct, since all initial design decisions and optimizations of the measurement tool served the goal of assessing the original construct, and the influence of foundational decisions might be hard to disperse without completely restarting the design process.

7. For instance, while the German "die Krankenschwester" ("the[female article] sister of the sick", i.e., nurse) have clear and stereotype-consistent grammatical genders, it is also possible for a word to have a neutral gender (grammatically), but a strong female/male gender bias.

4.3 Content validity

Content validity becomes relevant if we do not conceptualize model bias as unidimensional, but hypothesize the existence of subcomponents of the construct. In such cases, a bias measure usually involves the aggregation of subscores for these subcomponents (analogous to how different test scores are aggregated into one IQ score). For such composite scores, it would be important to establish content validity: the extent to which a measurement tool contains submeasures for all important subconstructs, without including construct-irrelevant content. If that comes to pass, we can make use of a whole library of psychometric literature and research methods (see e.g., factor analyses; Kline, 2014).

To take gender bias as an example, in human communication, different types of gender-based bias have been identified (see e.g., Zeinert et al., 2021; Stanczak & Augenstein, 2021). As language models are learning from natural language, an outstanding question would be whether these different types of bias are either directly represented in the model or whether they collectively (but not necessarily to the same extent) lead to a unidimensional manifestation of bias in the model. In Figure 4, we give an example of a possible subcategorization that may exist in a model's gender bias.

The existence of subconcepts of model bias has already been hinted at by some researchers (e.g., Du et al., 2021; Dev et al., 2022) and breaking down the bias construct into subcomponents and devising subtests for them comes with practical advantages: It is difficult to define "model bias" and a lack of a (consensus) definition hinders research on how to address it. Instead, it might be much easier to identify subcomponents of bias that most researchers do agree on and to develop (sub-)measures for them. If model bias was assessed by an aggregation of such submeasure scores, disagreements about the bias construct could be expressed by individual researchers' choices of submeasures to include in their aggregates.

We believe that discussing the subconcepts and different manifestations of gender bias will be important for the development of valid bias measures. Subconcepts of model bias might become especially relevant when considering other languages and bias types, since some manifestations may or may not be shared cross-culturally (see e.g., also our discussion of how gender is specified differently in different cultures in Section 5.2). However, identifying subconcepts may prove very difficult (e.g., techniques like factor analysis might statistically identify subcomponents of model bias for which we do not have intuitive explanations of what they mean), and these hypotheses should be thoroughly tested.

5. From theory to practice: Designing good bias measures

How do we put the lesson from psychometrics into practice when designing bias measures? In this section, we provide examples of practical questions to ask during three different stages of the process (Section 5.1). We then focus on validity concerns and other issues that may arise when generalizing English gender bias measures to other languages and bias types, and again provide some guidelines for ensuring the quality of bias measures in this particular setting (Section 5.2).

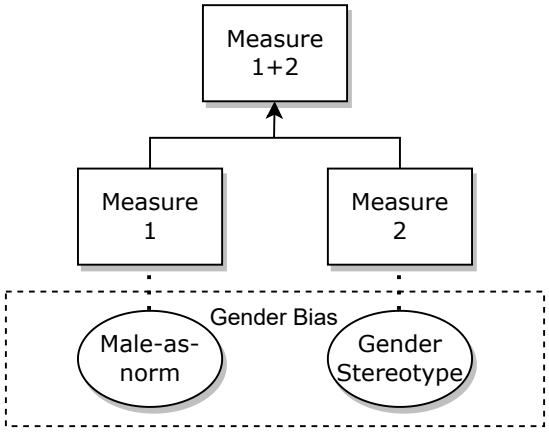

Figure 4: **Content validity:** in this example, *male-as-norm bias* and *gender stereotypes* are hypothesized to be separate subconcepts of the model's construct gender bias. The *male-as-norm* bias reflects the idea that the male gender is assumed as default, unless explicitly indicated otherwise (Danesi, 2014); This could also be reflected by a high prior for male pronouns. Gender stereotypes can refer to a broad category of phenomena where certain genders are associated with social norms, roles, or attributes and traits (e.g., women are seen as more passive; Eagly et al., 2020).

5.1 Guidelines for designing and evaluating bias measures

In the following, we discuss questions and considerations informed by psychometrics, as they apply to three different phases of the bias measure development cycle: *(I) the preparatory phase before designing the measure, (II) the evaluation phase of reliability and construct validity*, and *(III) the post-development phase in which results and limitations are communicated.* Our list of questions is not intended to be exhaustive; it just provides some examples for the types of issues researchers should consider when developing such a measure. In that, it should be seen as complementing other guidelines from the literature (e.g., Blodgett et al., 2020, 2021; Dev et al., 2022; Talat et al., 2022).

(I) Preparation phase: Understanding the task and sociotechnical context

- What are relevant technical details about the model (e.g., is it an autoregressive or masked LM)? What do we know about the training data? Which (downstream) tasks are being considered for this model?

- What are the relevant forms of bias to measure? For what language(s) and cultural context(s) do we need to apply the measures? To what extent can we reuse existing measures designed for other bias types (e.g., US-centric binary gender bias)? What can we learn about the bias from the social-scientific literature, stakeholders, and potential end-users?

- Given this information, how can the bias be operationalized, while avoiding pitfalls known in the literature? How can these decisions and assumptions (as well as the decision-making process) be documented to enable later transparency?

(II) Evaluation phase: Assessing the reliability and construct validity

- What are the constraints and available resources for conducting a reliability and validity assessment (e.g., computational resources, access to training data, access to the internal states of the model)?

- What reliability tests from Section 3 are relevant and feasible (e.g., is the *internal consistency* of the measure relevant)? Can we test the reliability of the LM's representations that are used to develop the bias measure? Is it computationally feasible to retrain the model to assess the bias measure's *test-retest reliability*? Can we assess the extent to which the bias measure generalizes to a different language model (architecture)?

- What validity tests from Section 4 are feasible (e.g., for testing *convergent validity*, are there other relevant bias measures, behavioral data, and downstream tasks to compare our bias measures with)? Insofar as relevant data is not available, how feasible would it be to obtain (e.g., can we obtain behavioral experimental or survey data of stakeholders, for assessing perceived bias/harms in downstream behavior)? Are there relevant measures for testing the *divergent validity*? Could my measure accommodate subcategorizations of bias; if so, how do I assess my measure's *content validity*?

(III) Post-development phase: Communicating the results and limitations

- For what particular sociotechnical context have the measures been validated (e.g., autoregressive language models trained on Dutch news articles)? For what contexts does the assessment hold, and when do we need to perform a new reliability and validity assessment (see Section 5.2)? Assuming we reached acceptable levels of reliability and validity, what limitations of the bias measure must still be communicated to stakeholders?

- Considering that bias measures are imperfect, how do these imperfections affect the decisions that can be made based on the measure and/or tested model? When is a model sufficiently "unbiased" for it to be used in a particular context? Will the model be re-evaluated in the future, if more powerful (or more theoretically solid) bias measures are developed?

- How does the cultural context influence the undesirability of a bias? How may the definition of undesirable bias change depending on a different cultural context? What would change if your assumptions about the conceptualization of bias might prove wrong or outdated later?

5.2 Operationalization and validity issues when generalizing bias measures

As our last set of questions (III) implies, we believe that bias measures, validated for one particular language or cultural context, do not necessarily transfer well to a different context. This is particularly problematic, considering that most research on bias in NLP is focused on one type of bias in one language: gender bias for the English language (Field et al., 2021; Talat et al., 2022). Indeed, the bias evaluation of NLP technologies in the multilingual and multicultural setting is especially prone to validity issues (Talat et al., 2022; Blodgett et al., 2021; Malik et al., 2022) and bias mitigation efforts do not necessarily transfer between languages even within the same multilingual model (Gonen et al., 2022) or between types of biases (Jin et al., 2021).

As an illustration of the difficulties of generalizing a measure beyond its initial purpose, we outline three types of obstacles to generalizing English gender bias measures to other languages and bias types. Such generalization attempts might (1) fail at the operationalization level due to linguistic properties of the target language, (2) involve insufficient modeling of ethnic backgrounds or (3) yield invalid/inconclusive bias scores, since marginalized identities are underrepresented in the training data.

Differing linguistic properties of the target language English gender bias detection relies heavily on how and where the English language marks gender as a grammatical feature. For instance, English marks gender explicitly only on personal and possessive pronouns (Audring, 2016), but not for example on nouns such as profession words. Some bias measures exploit this fact (e.g., Rudinger et al., 2018; Zhao et al., 2018). However, gender systems vary considerably across languages (Audring, 2016). Hence, operationalization of bias measures such as used by e.g., Rudinger et al., Zhao et al., might not be very meaningful in another languages if their transfer involves simple translations of word lists. See Table 2 for examples.

Insufficient modelling of ethnic backgrounds For the assessment of gender bias, gender is usually modelled as a binary category with gendered nouns, pronouns and names assigned to either category for the sake of comparison. Extending this approach to racial bias is non-trivial, since many ethnic groups exist and power dynamics are particular to any one cultural context (Blodgett et al., 2020). Hence, choosing demographic categories for a comparison poses difficult modelling choices. Choosing correct proxies for demographic backgrounds (e.g., names) is another challenge at the operationalization level (Sweeney, 2013; Fryer Jr & Levitt, 2004; Wood-Doughty et al., 2018). Careful thought is required to arrive at a good selection of proxies. We summarize concerns around such modelling choices for bias measurements in Table 3.

Underrepresentation of marginalized identities Even if a English gender bias measures can be adapted for other contexts, problems can ensue. Protected identities other than gender are linguistically less explicitly marked (e.g., in English, there are no pronouns to convey a person's age or sexual orientation). Hence, bias measures often involve direct mentions of an identity (e.g., "the Muslim boy") or names and dialects that are linked to cultural backgrounds (e.g., "Tyrone" vs. "John" in a US context). Since such identities are underrepresented in the data, comparing model behavior on different identities might not be conclusive. Hence, the resulting bias measure could be partially invalid on a technical

Obstacles to Applicability	Example
Translating gendered word pairs (e.g., *he-she*, *his-hers*) from (Bolukbasi et al., 2016).	In Korean and Hungarian *he/she* are gender-neutral and in German, *she* can also signify *they*, *them* or *you* (Cho et al., 2019; Prates et al., 2020; Chen et al., 2021).
Translating WEAT lists (Caliskan et al., 2017) containing science words (e.g., *science*, *technology*, *physics*, *NASA*). and arts words (e.g., *poetry*, *art*, *Shakespeare*,*dance*.)	In Hindi, no accurate translations of *Shakespeare* and *NASA* exist (Malik et al., 2022).
Gendering of profession words (e.g., *cashier*, *developer*, *waitress*) used in (e.g., Zhao et al., 2018)	Nouns can be gendered or neutral, e.g., Italian: *scultore* (m)/*scultrice* (f) (English: *sculptor*) vs. *fiorista* (m/f) (English: *florist*). Similarly for Spanish, French, German (Zhou et al., 2019; Chen et al., 2021).

Table 2: Obstacles to operationalizing bias measures for other languages when applying existing English gender bias measures to other languages.

level. We summarize concerns around validity for both of the aforementioned approaches in Table (4). By extension, under-representation concerns also apply to bias measures designed for intersections of biases (Wolfe & Caliskan, 2021; Tan & Celis, 2019; Kiritchenko & Mohammad, 2018; Kirk et al., 2021).

Overall, when one adapts an existing bias measure to a new language, type of bias or merely applies it in a different cultural context, one should keep in mind that existing validity and reliability results do not necessarily carry over to the new context. Rather, to ensure its validity, a measure needs to be reassessed whenever it is applied in a new context. Conversely, when approaching bias detection with English gender bias detection methods in mind, researchers might miss out on crucial manifestations of bias (Ciora et al., 2021).[8]

We end this section with a list of questions researchers could ask themselves before generalizing existing bias measures to new cultural contexts.

8. For example, in Turkish, gender markings of nouns are optional and bias might show itself in whether or not gender is explicitly marked. For instance, to translate the words sister/brother into Turkish, there exists only one gender-neutral translation 'sibling' which is optionally accompanied by a word for female/male. When translating "My sister/brother is a soccer player" into Turkish, Google Translate explicitly marks the gender in the former case but not in the latter (Ciora et al., 2021). Another example is Chinese, where words combine the meaning of their constituent characters and model training is based on character information. Bias detection methods need to take this into account to not miss other such manifestations of bias (Jiao & Luo, 2021).

Validity Concern	Example
Choice of racial categories	Racial categories do not generalize across cultural contexts (Ghosh et al., 2021; Malik et al., 2022; Abid et al., 2021). Racial categories are socially construed (Hanna et al., 2020).
Non-binary persons as a third category	Modelling non-binary persons as a homogeneous third class does not reflect the fluidity of gender (Dev & Monajatipoor, 2021).
Point of comparison for non-binary categories	Comparing model behavior on any pair of groups, such as *Asian* vs. *Latino* might not be meaningful. Comparison between *African American* and *White American*, while reflecting existing power dynamics, might entrench the dominant group as the default point of comparison (Blodgett et al., 2020).
Names lists as proxy for race	Most lists (e.g., Caliskan et al., 2017; Garg et al., 2018) are based on potentially outdated ratings of *White American* and *African American* names by psychology students in 1998 (Greenwald et al., 1998; Sweeney, 2013; Fryer Jr & Levitt, 2004; Wood-Doughty et al., 2018).

Table 3: Pitfalls in defining demographic attributes for bias measures.

(IV) Generalizing bias measures to new cultural contexts

- For which target language, cultural context and type of bias was the bias measure developed? How is membership of a potentially marginalized group marked linguistically, and how can these markers be used to operationalize a meaningful bias measure? How is gender encoded differently in the target language in comparison to English?
- Which social groups exist in the stakeholder's cultural context? What are their power dynamics? Which groups should be compared?
- Can we analyze to what extent linguistic markers of marginalized groups (e.g., names, pronouns, identity mentions) are represented or underrepresented in the training corpus? How might this affect bias scores?

6. Related work

Viewing bias measures as operationalizations of a hidden construct, such as model bias, is currently an uncommon approach in the field, as are systematic assessments of their construct validity and reliability. However, we are not the first to discuss validity and reliability concerns of existing bias measures.

Validity Concern	Example
Comparison of model behavior on names	*Female* and *African American* names are more often multiply-tokenized (i.e., mapped to multiple tokens) and their embeddings more often overfit to a small set of contexts (Wolfe & Caliskan, 2021).
Comparison of model behavior on identity mentions	Linguistically, dominant groups are usually not explicitly mentioned (e.g., LMs will have seen fewer examples of *able-bodied man* or *heterosexual man* than *disabled* or *gay man*. (Bucholtz & Hall, 2004; Blodgett et al., 2021; Kitzinger, 2005; Rendle-Short, 2005).
Comparison of he/she with neopronouns	*Ze*, *xe* and singular *they* are underrepresented in training corpora, so LM embeddings tend to be unstable (Dev & Monajatipoor, 2021).

Table 4: Names, pronouns and identity mentions that are associated with marginalized identities occur less often in the training data. Hence, contrasting model behavior on names, pronouns and identity mentions that function as a proxy for different identity groups, might compromise the validity of a bias measure.

In their survey of bias research in NLP, Blodgett et al. (2020) concluded that what researchers meant with *bias* was often poorly defined and inconsistent with the pronounced research goals of the field. The authors argued for more transparency about what the researchers were actually trying to measure/mitigate and proposed that researchers explicitly ground bias measures in the downstream harms of NLP systems (as we also discuss in Section 4.1). Another survey by Blodgett et al. (2021) — but of measurement tools based on constrastive sets such as CrowS-Pairs (Nangia et al., 2020) and StereoSet (Nadeem et al., 2021) — categorized an extensive set of examples of bad operationalizations, which threaten the construct validity of these bias measurement benchmarks. Dev et al. (2022) propose a comprehensive set of questions for improving the documentation of bias measures, including questions concerning the validity (although not mentioning the term explicitly).

While papers like these point towards issues in the field that we tried to help addressing here, few works have, like us, explicitly applied concepts of psychometrics to the issue of measuring bias (noteworthy exceptions being Zhang et al., 2020; Du et al., 2021). Du et al. (2021) conduct an extensive evaluation of the *reliability* of various static word embedding bias measures. Their work is a good example of the types of reliability evaluations that we advocate for in Section 3.

In the related field of algorithmic fairness, some works have made explicit the distinction between construct and operationalizations (e.g., Friedler et al., 2021; Jacobs & Wallach, 2021). Perhaps closest to our work, is the one of Jacobs and Wallach (2021), who, similar to our paper, introduce key concepts from psychometrics, including a discussion of types of reliability and construct validity that could be relevant for computational scientists. However, their focus on measuring fairness in algorithmic decision-making differs from our

focus on measuring model bias — a related, but different construct with other measurement tools. As a result, we discuss different methodologies, open questions, and arrive at other recommendations.

7. Conclusions

Algorithmic bias in NLP is an inherently complex phenomenon due to its sociotechnical and context-sensitive nature (Blodgett et al., 2020; Talat et al., 2022). As a result, researchers face many challenges in the development of measuring and mitigation tools. In this paper, we addressed the question of how we can test the quality of bias measures, despite these complexities. In our view, part of the answer is to adopt useful psychometric vocabulary (and methodology) when discussing the topic of bias. Psychological measurements share some of the same challenges as NLP bias (e.g., unobservability of the construct, disagreements between researchers about what ought to be measured). Consequently, their ways of addressing these challenges (e.g., frameworks for assessing reliability and validity) might prove valuable to NLP, as well.

Of course, adopting a psychometric framework will not solve all issues. Beyond psychometricians, there is a great need for involving other experts (e.g., social scientists, psychologists, philosophers, and ethicists) and stakeholders (e.g., designers, owners, and users of these NLP systems, and those potentially harmed by its implementation) in the measurement tool design process (see also Blodgett et al., 2020; Bender et al., 2021; Kiritchenko et al., 2021; Talat et al., 2022; Dev et al., 2022, i.a.). That is firstly because designing good measurement tools requires a thorough understanding of the sociocultural context of the (model) bias (also whenever it is applied to new settings; see Section 5) — this necessitates other expertise and life experiences than psychometrics. Secondly, the use of a psychometric lens has its limits. While we have access to a rich history of methodologies and insights from the field, not all (readily) apply to the NLP setting. For instance, methodologies designed for human test subjects may be intractable for language models (e.g., because we need too many "test-takers"), or because the analogy between a model and a person breaks down (e.g., a language model is not subject to time in the same way as people are). As a result, to address all outstanding questions, a multidisciplinary approach will be required.

Indirectly, framing the discussion of bias through psychometric terms may aid such multidisciplinary collaborations and involvements of stakeholders. Subjective/normative choices must be made about where the responsibility of an NLP practitioner ends and where other experts or stakeholders should be involved (in other words, discourse about questions like "When is a question sufficiently normative to necessitate stakeholder-involvement?"). A first step towards identifying questions that stakeholders should weigh in on could be to identify disagreements that currently exist within the field of bias measurement — especially those that do not have empirical answers. Such disagreements can, however, only be unearthed if researchers are explicit about the assumptions they make. We hope that our discussion of different types of assumptions (e.g., about construct or operationalization) will help NLP researchers refine and communicate their individual understandings of bias — mitigating the current conceptual confusion (Blodgett et al., 2020; Dev et al., 2022).

Beyond transparency, we believe that utilizing the psychometric vocabulary we emphasized in this paper could also improve the rate of progress in the NLP bias field. Specifically,

we hope our advice can help improve the communication between researchers by helping to contextualize findings (e.g., as pertaining to a particular operationalization versus a particular construct) and by specifying possible points of theoretical convergence and divergence. Beyond aiding communication, efforts towards evaluating the validity and reliability of contemporary (and future) bias measures can also contribute to this increased efficiency of the bias field: If we default to assuming that a measure is reliable and valid, we risk wasting time and resources on the wrong targets for improvement efforts (e.g., not focusing on reliability improvements for an unreliable measure) or on bias measures that the field should have moved on from (if their validity and reliability issues are insurmountable).

Like Jacobs and Wallach (2021), we are only an early effort towards applying measurement theory and psychometric concepts to AI. As such, we do not want to imply that our perspectives on the topic are definite or gospel. Instead, we hope that we further opened the door towards applying psychometric concepts to AI and invite theoretical discussions of their merits (or conversely, inapplicability) to NLP bias research.

Acknowledgments

This publication is part of the project "The biased reality of online media - Using stereotypes to make media manipulation visible" (with project number 406.DI.19.059) of the research programme Open Competition Digitalisation-SSH, which is financed by the Dutch Research Council (NWO).

References

Abid, A., Farooqi, M., & Zou, J. (2021). Large language models associate Muslims with violence. *Nature Machine Intelligence*, *3*(6), 461–463. Number: 6 Publisher: Nature Publishing Group.

Amidei, J., Piwek, P., & Willis, A. (2020). Identifying Annotator Bias: A new IRT-based method for bias identification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4787–4797, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Audring, J. (2016). Gender..

Balayn, A., & Gürses, S. (2021). Beyond debiasing: Regulating ai and its inequalities..

Barocas, S., Crawford, K., Shapiro, A., & Wallach, H. (2017). The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual conference of the special interest group for computing, information and society*.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning*. fairmlbook.org. `http://www.fairmlbook.org`.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 610–623, New York, NY, USA. Association for Computing Machinery.

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, Online. Association for Computational Linguistics.

Blodgett, S. L., Lopez, G., Olteanu, A., Sim, R., & Wallach, H. (2021). Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1004–1015, Online. Association for Computational Linguistics.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Bommasani, R., et al. (2022). On the Opportunities and Risks of Foundation Models..

Bordia, S., & Bowman, S. R. (2019). Identifying and Reducing Gender Bias in Word-Level Language Models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pp. 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity.. *Psychological review*, *111*(4), 1061.

Bucholtz, M., & Hall, K. (2004). Language and identity. *A companion to linguistic anthropology*, *1*, 369–394.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Cao, Y., Pruksachatkun, Y., Chang, K.-W., Gupta, R., Kumar, V., Dhamala, J., & Galstyan, A. (2022a). On the Intrinsic and Extrinsic Fairness Evaluation Metrics for Contextualized Language Representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 561–570, Dublin, Ireland. Association for Computational Linguistics.

Cao, Y., Sotnikova, A., Daumé III, H., Rudinger, R., & Zou, L. (2022b). Theory-Grounded Measurement of U.S. Social Stereotypes in English Language Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1276–1295, Seattle, United States. Association for Computational Linguistics.

Chen, Y., Mahoney, C., Grasso, I., Wali, E., Matthews, A., Middleton, T., Njie, M., & Matthews, J. (2021). Gender Bias and Under-Representation in Natural Language Processing Across Human Languages. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pp. 24–34, New York, NY, USA. Association for Computing Machinery.

Cheng, L., Varshney, K. R., & Liu, H. (2021). Socially Responsible AI Algorithms: Issues, Purposes, and Challenges. *Journal of Artificial Intelligence Research*, *71*, 1137–1181.

Cho, W. I., Kim, J. W., Kim, S. M., & Kim, N. S. (2019). On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pp. 173–181.

Ciora, C., Iren, N., & Alikhani, M. (2021). Examining covert gender bias: A case study in turkish and english machine translation models. In *Proceedings of the 14th International Conference on Natural Language Generation*, pp. 55–63.

Crawford, K. (2017). The trouble with bias.. Keynote at Neural Information Processing Systems (NIPS'17).

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., Hormozdiari, F., Houlsby, N., Hou, S., Jerfel, G., Karthikesalingam, A., Lucic, M., Ma, Y., McLean, C., Mincu, D., Mitani, A., Montanari, A., Nado, Z., Natarajan, V., Nielson, C., Osborne, T. F., Raman, R., Ramasamy, K., Sayres, R., Schrouff, J., Seneviratne, M., Sequeira, S., Suresh, H., Veitch, V., Vladymyrov, M., Wang, X., Webster, K., Yadlowsky, S., Yun, T., Zhai, X., & Sculley, D. (2022). Underspecification Presents Challenges for Credibility in Modern Machine Learning. *Journal of Machine Learning Research*, *23*(226), 1–61.

Danesi, M. (2014). *Dictionary of media and communications*. Routledge.

Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pp. 4691–4697, Melbourne, Australia. AAAI Press.

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., & Kalai, A. T. (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*, FAT* '19, pp. 120–128, New York, NY, USA. Association for Computing Machinery.

De Cao, N., Schmid, L., Hupkes, D., & Titov, I. (2021). Sparse Interventions in Language Models with Differentiable Masking.. arXiv:2112.06837 [cs].

Delobelle, P., Tokpo, E., Calders, T., & Berendt, B. (2022). Measuring Fairness with Biased Rulers: A Comparative Study on Bias Metrics for Pre-trained Language Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1693–1706, Seattle, United States. Association for Computational Linguistics.

Dev, S., & Monajatipoor, M. (2021). Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Dev, S., Sheng, E., Zhao, J., Amstutz, A., Sun, J., Hou, Y., Sanseverino, M., Kim, J., Nishi, A., Peng, N., & Chang, K.-W. (2022). On Measures of Biases and Harms in NLP.. arXiv:2108.03362 [cs].

Dinan, E., Fan, A., Wu, L., Weston, J., Kiela, D., & Williams, A. (2020). Multi-Dimensional Gender Bias Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 314–331, Online. Association for Computational Linguistics.

Du, Y., Fang, Q., & Nguyen, D. (2021). Assessing the Reliability of Word Embedding Gender Bias Measures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10012–10034, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eagly, A. H., Nater, C., Miller, D. I., Kaufmann, M., & Sczesny, S. (2020). Gender stereotypes have changed: A cross-temporal meta-analysis of U.S. public opinion polls from 1946 to 2018. *American Psychologist*, *75*, 301–315. place: US publisher: American Psychological Association.

Ethayarajh, K. (2020). Is Your Classifier Actually Biased? Measuring Fairness under Uncertainty with Bernstein Bounds. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2914–2919, Online. Association for Computational Linguistics.

Ethayarajh, K., Duvenaud, D., & Hirst, G. (2019). Understanding Undesirable Word Embedding Associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1696–1705, Florence, Italy. Association for Computational Linguistics.

Field, A., Blodgett, S. L., Waseem, Z., & Tsvetkov, Y. (2021). A Survey of Race, Racism, and Anti-Racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1905–1925, Online. Association for Computational Linguistics.

Field, A., & Tsvetkov, Y. (2020). Unsupervised Discovery of Implicit Gender Bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 596–608, Online. Association for Computational Linguistics.

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, *12*(1). tex.ids= founta2018LargeScaleCrowdsourcinga number: 1.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2021). The (Im)possibility of fairness: different value systems require different mechanisms for fair decision making. *Communications of the ACM*, *64*(4), 136–143.

Fryer Jr, R. G., & Levitt, S. D. (2004). The causes and consequences of distinctively black names. *The Quarterly Journal of Economics*, *119*(3), 767–805.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644. Publisher: Proceedings of the National Academy of Sciences.

Ghosh, S., Baker, D., Jurgens, D., & Prabhakaran, V. (2021). Cross-geographic bias detection in toxicity modeling..

Goldfarb-Tarrant, S., Marchant, R., Muñoz Sánchez, R., Pandya, M., & Lopez, A. (2021). Intrinsic Bias Metrics Do Not Correlate with Application Bias. In *Proceedings of*

the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 1926–1940, Online. Association for Computational Linguistics.

Gonen, H., & Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Gonen, H., Ravfogel, S., & Goldberg, Y. (2022). Analyzing Gender Representation in Multilingual Models. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pp. 67–77, Dublin, Ireland. Association for Computational Linguistics.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test.. *Journal of personality and social psychology*, *74*(6), 1464.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*(1), 17–41. Place: US Publisher: American Psychological Association.

Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*. Springer Science & Business Media.

Hanna, A., Denton, E., Smart, A., & Smith-Loud, J. (2020). Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT∗'20, p. 501–512, New York, NY, USA. Association for Computing Machinery.

Hogenboom, S. A. M., Schulz, K., & Van Maanen, L. Implicit association tests: Stimuli validation from participant responses.. In principle accepted.

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, *15*(8), e12432.

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020). Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5491–5501, Online. Association for Computational Linguistics.

Jacobs, A. Z., & Wallach, H. (2021). Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 375–385, New York, NY, USA. Association for Computing Machinery.

Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, Online. Association for Computational Linguistics.

Jiao, M., & Luo, Z. (2021). Gender Bias Hidden Behind Chinese Word Embeddings: The Case of Chinese Adjectives. In *Proceedings of the 3rd Workshop on Gender Bias*

in Natural Language Processing, pp. 8–15, Online. Association for Computational Linguistics.

Jin, X., Barbieri, F., Kennedy, B., Mostafazadeh Davani, A., Neves, L., & Ren, X. (2021). On Transferability of Bias Mitigation Effects in Language Model Fine-Tuning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3770–3783, Online. Association for Computational Linguistics.

Kiritchenko, S., & Mohammad, S. (2018). Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pp. 43–53. Association for Computational Linguistics.

Kiritchenko, S., Nejadgholi, I., & Fraser, K. C. (2021). Confronting Abusive Language Online: A Survey from the Ethical and Human Rights Perspective. *Journal of Artificial Intelligence Research*, *71*, 431–478.

Kirk, H. R., Jun, Y., Volpin, F., Iqbal, H., Benussi, E., Dreyer, F., Shtedritski, A., & Asano, Y. (2021). Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models. In *Advances in Neural Information Processing Systems*, Vol. 34, pp. 2611–2624. Curran Associates, Inc.

Kitzinger, C. (2005). "Speaking as a Heterosexual": (How) Does Sexuality Matter for Talk-in-Interaction?. *Research on Language and Social Interaction*, *38*(3), 221–265. Publisher: Routledge.

Kline, P. (2014). *An easy guide to factor analysis*. Routledge.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & quantity*, *47*(4), 2025–2047.

Kurita, K., Vyas, N., Pareek, A., Black, A. W., & Tsvetkov, Y. (2019). Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pp. 166–172, Florence, Italy. Association for Computational Linguistics.

Lalor, J. P., Wu, H., & Yu, H. (2016). Building an evaluation scale using item response theory. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, Vol. 2016, p. 648. NIH Public Access.

Limisiewicz, T., & Mareček, D. (2022). Don't Forget About Pronouns: Removing Gender Bias in Language Models Without Losing Factual Gender Information. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pp. 17–29, Seattle, Washington. Association for Computational Linguistics.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing.. arXiv:2107.13586 [cs].

Malik, V., Dev, S., Nishi, A., Peng, N., & Chang, K.-W. (2022). Socially Aware Bias Measurements for Hindi Language Representations.. arXiv:2110.07871 [cs].

Meade, N., Poole-Dayan, E., & Reddy, S. (2022). An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

Montuschi, P., Gatteschi, V., Lamberti, F., Sanna, A., & Demartini, C. (2013). Job recruitment and job seeking processes: how technology can help. *It professional*, *16*(5), 41–49.

Nadeem, M., Bethke, A., & Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5356–5371, Online. Association for Computational Linguistics.

Nangia, N., Vania, C., Bhalerao, R., & Bowman, S. R. (2020). CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1953–1967, Online. Association for Computational Linguistics.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., Ishiyama, J., Karlan, D., Kraut, A., Lupia, A., Mabry, P., Madon, T., Malhotra, N., Mayo-Wilson, E., McNutt, M., Miguel, E., Paluck, E. L., Simonsohn, U., Soderberg, C., Spellman, B. A., Turitto, J., VandenBos, G., Vazire, S., Wagenmakers, E. J., Wilson, R., & Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. Publisher: American Association for the Advancement of Science.

Névéol, A., Dupont, Y., Bezançon, J., & Fort, K. (2022). French CrowS-Pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8521–8531, Dublin, Ireland. Association for Computational Linguistics.

Prates, M. O., Avelar, P. H., & Lamb, L. C. (2020). Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, *32*(10), 6363–6381.

Rendle-Short, J. (2005). 'i've got a paper-shuffler for a husband': indexing sexuality on talk-back radio. *Discourse & Society*, *16*(4), 561–578.

Rudinger, R., Naradowsky, J., Leonard, B., & Van Durme, B. (2018). Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Scao, T. L., et al. (2022). BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.. arXiv:2211.05100 [cs].

Schick, T., Udupa, S., & Schütze, H. (2021). Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, *9*, 1408–1424.

Shah, D. S., Schwartz, H. A., & Hovy, D. (2020). Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5248–5264, Online. Association for Computational Linguistics.

Stanczak, K., & Augenstein, I. (2021). A Survey on Gender Bias in Natural Language Processing.. arXiv:2112.14168 [cs].

Stanovsky, G., Smith, N. A., & Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1679–1684.

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., & Wang, W. Y. (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1630–1640, Florence, Italy. Association for Computational Linguistics.

Sweeney, L. (2013). Discrimination in online ad delivery: Google ads, black names and white names, racial discrimination, and click advertising. *Queue*, *11*(3), 10–29.

Talat, Z., Névéol, A., Biderman, S., Clinciu, M., Dey, M., Longpre, S., Luccioni, S., Masoud, M., Mitchell, M., Radev, D., Sharma, S., Subramonian, A., Tae, J., Tan, S., Tunuguntla, D., & van der Wal, O. (2022). You reap what you sow: On the Challenges of Bias Evaluation Under Multilingual Settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pp. 26–41, virtual+Dublin. Association for Computational Linguistics.

Tan, Y. C., & Celis, L. E. (2019). Assessing Social and Intersectional Biases in Contextualized Word Representations. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc.

Tontodimamma, A., Fontanella, L., Anzani, S., & Basile, V. (2022). An Italian lexical resource for incivility detection in online discourses. *Quality & Quantity*, *56*.

Verma, S., Vieweg, S., Corvey, W., Palen, L., Martin, J., Palmer, M., Schram, A., & Anderson, K. (2011). Natural Language Processing to the Rescue? Extracting "Situational Awareness" Tweets During Mass Emergency. *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(1), 385–392. Number: 1.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., & Shieber, S. (2020). Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In *Advances in Neural Information Processing Systems*, Vol. 33, pp. 12388–12401. Curran Associates, Inc.

Wagner, C., Graells-Garrido, E., Garcia, D., & Menczer, F. (2016). Women through the glass ceiling: gender asymmetries in Wikipedia. *EPJ Data Science*, *5*(1), 5.

van der Wal, O., Jumelet, J., Schulz, K., & Zuidema, W. (2022). The Birth of Bias: A case study on the evolution of gender bias in an English language model. In *Proceedings*

of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pp. 75–75, Seattle, Washington. Association for Computational Linguistics.

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., et al. (2018). Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, *77*, 34–49.

Way, A. (2018). Quality expectations of machine translation. In *Translation quality assessment*, pp. 159–178. Springer.

Webster, K., Wang, X., Tenney, I., Beutel, A., Pitler, E., Pavlick, E., Chen, J., Chi, E., & Petrov, S. (2021). Measuring and Reducing Gendered Correlations in Pre-trained Models.. arXiv:2010.06032 [cs].

Weinberg, L. (2022). Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches. *Journal of Artificial Intelligence Research*, *74*, 75–109.

Whitlock, M., & Schluter, D. (2015). *The analysis of biological data*, Vol. 768. Roberts Publishers.

Wolfe, R., & Caliskan, A. (2021). Low Frequency Names Exhibit Bias and Overfitting in Contextualizing Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 518–532, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wong, K., & Paritosh, P. (2022). k-rater reliability: The correct unit of reliability for aggregated human annotations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 378–384.

Wood-Doughty, Z., Andrews, N., Marvin, R., & Dredze, M. (2018). Predicting twitter user demographics from names alone. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pp. 105–111.

Zeinert, P., Inie, N., & Derczynski, L. (2021). Annotating Online Misogyny. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3181–3197, Online. Association for Computational Linguistics.

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

Zhang, Y., Zhang, Y., Halpern, B. M., Patel, T., & Scharenborg, O. (2022). Mitigating bias against non-native accents. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Vol. 2022, pp. 3168–3172.

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., & Chang, K.-W. (2019). Gender Bias in Contextualized Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, Volume 1 (Long and Short Papers), pp. 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., & Sun, M. (2020). How does nlp benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5218–5230.

Zhou, P., Shi, W., Zhao, J., Huang, K.-H., Chen, M., Cotterell, R., & Chang, K.-W. (2019). Examining gender bias in languages with grammatical gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5276–5284.