UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Moving beyond harm

A critical review of how NLP research approaches discrimination

Schulz, K.; Lanzing, M.; Martinez Brenner, G.

DOI
10.48550/arXiv.2508.04504

Publication date
2025

Document Version
Final published version

License
CC BY-NC-SA

Link to publication

Citation for published version (APA):
Schulz, K., Lanzing, M., & Martinez Brenner, G. (2025). *Moving beyond harm: A critical review of how NLP research approaches discrimination*. (v1 ed.) ArXiv. https://doi.org/10.48550/arXiv.2508.04504

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:13 Nov 2026

arXiv:2508.04504v1 [cs.CY] 6 Aug 2025

Moving beyond harm. A critical review of how NLP research approaches discrimination

Katrin Schulz, ILLC, University of Amsterdam
Marjolein Lanzing, Philosophy Department, University of Amsterdam
Giulia Martinez Brenner, University of Amsterdam

2025
January

Abstract

How to avoid discrimination in the context of NLP technology is one of the major challenges in the field. We propose that a different and more substantiated framing of the problem could help to find more productive approaches. In the first part of the paper we report on a case study: a qualitative review on papers published in the ACL anthologies 2022 on discriminatory behavior of NLP systems. We find that the field (i) still has a strong focus on a technological fix of algorithmic discrimination, and (ii) is struggling with a firm grounding of their ethical or normative vocabulary. Furthermore, this vocabulary is very limited, focusing mostly on the terms "harm" and "bias". In the second part of the paper we argue that addressing the latter problems might help with the former. The understanding of algorithmic discrimination as a technological problem is reflected in and reproduced by the vocabulary in use. The notions of "harm" and "bias" implicate a narrow framing of the issue of discrimination as one of the system-user interface. We argue that instead of "harm" the debate should make "injustice" the key notion. This would force us to understand the problem of algorithmic discrimination as a systemic problem. Thereby, it would broaden our perspective on the complex interactions that make NLP technology participate in discrimination. With that gain in perspective we can consider new angles for solutions.

1 Introduction

In recent years a lot of attention has been paid to machine learning applications that show unfair or biased behavior towards particular groups or individuals. This

led to growing societal concerns about harmful discrimination and reproduction of structural inequalities once these technologies become institutionalized in society. Identifying and resolving algorithmic bias and achieving algorithmic fairness has therefore become an important research field. With the rise of large language models (LLMs) and the speed with which they are integrated in all aspects of life, this has been particularly true in the area of Natural Language Processing (NLP).

However, it has been argued that in AI research and policy, the remedies against discriminatory algorithms are often narrowly framed as addressing technological design problems aimed at compliance, rather than complex, structural, social-political issues. This leads to a highly techno-centric and technocratic approach towards algorithmic fairness Balayn and Gurses (2021). Furthermore, it has been argued that this framing stands in the way of successfully addressing the problem of negative societal impact Blodgett et al. (2020); Birhane (2021).

In this paper we want to engage with this line of argument. Firstly, we will investigate whether and in what sense it still holds that algorithmic discrimination is approached in a techno-centric way. To this end, we perform a critical interpretative literature review, drawing a sample from the publications in the ACL anthologies for the year 2022 (see Section 2 for a description of the methodology). We find that nearly all the work on the topic of algorithmic discrimination in our sample still strongly focuses on finding technological solutions for the problem (see Subsection 3.3). We also observe that this comes together with a rather narrow framing of the mechanisms underlying the problem (Subsection 3.2) and a weak normative/ethical grounding of the problem addressed (Subsection 3.1).

Building on these results, in Section 5 we provide a proposal for how to overcome these limitations. The driving idea of this proposal is that in order to see the full range of factors responsible for algorithmic discrimination, and, thus, move beyond a technological solution, we need a thorough understanding of the problem. The proposal, therefore, targets the particular way of how the field views and conceptualizes the problem of algorithmic discrimination. Concretely, we will criticize the language used to develop the normative grounding of NLP research on algorithmic discrimination. As our case study shows, the current debate centers around the notions of *bias* and *harm*. We will argue that (i) *bias* in itself cannot provide ethical grounding and (ii) *harm* is neither sufficient nor necessary for discrimination to take place.

The alternative framing that we propose centers around the notions of *structural and systemic injustice* Haslanger (2022, 2023, 2024). Focusing on structural and systemic injustice instead of individual harms creates space to consider more complex mechanisms underlying patterns of algorithmic discrimination. It also changes the type of mechanisms we are looking for: a structural and systematic injustice approach would reveal unjust social-political structures connected to ide-

ologies of oppression as opposed to individual actions of (AI) agents or particular design choices made in the development of this technology. We hope that such a re-framing of the problem will help to create awareness among researchers and engineers about alternative ways to approach developing responsible NLP technology, but will also lead to a broader engagement of the academic and non-academic community in the discussion of how to address ethical issues of NLP technology.

Engaging with ethical questions can be complex and challenging. Before we continue with the main body of the paper, let us be explicit about what kind of unethical behavior of NLP technology we are targeting here. By restricting our discussion to algorithmic discrimination, we are looking specifically at the involvement of language technology in practices that lead to the unjust exclusion or mistreatment of groups or individuals. This still includes a broad spectrum of ways in which NLP technology can be involved in unethical practices, but not necessarily all such practices. It does include problems often discussed using notions like discrimination, fairness and bias, among others.

Second, making claims about unjust practices involves taking a normative stance. There is a lot of discussion about what morally wrong behavior is and we do not want to build this work on one very specific approach to ethics and, thereby, limit the reach of our proposal. However, drawing some bottom line, making some basic assumptions about what we value is unavoidable when grounding a normative position. For this paper our baseline is democracy. This will be the foundation on which we build our normative claims and arguments. There is a very lively debate in political science and philosophy about what democracy exactly entails and how it should be implemented.[1] To make our argumentation as broadly applicable as possible, we choose here a very general and thin concept of democracy. Let us agree that this democracy encompasses a liberal state of justice Rawls (2001). This includes valuing individual (human) rights, without failing to take into account the importance to recognize collectives and differences in order to grasp structural and systemic injustices that escape the individualist eye Young (1990). The underlying democratic ideal is that all members and dwellers of a democratic society should have (access to) equal rights and equal opportunities to participate, including the basic conditions necessary for this Rawls (2001). Building on this, the normative question of what we should expect of technology – which lies at the core of this paper – comes down to asking whether technology supports or hinders our democratic values and the processes that we have put in place to foster equality and freedom for all members of society. First of all, this grounds the normative stance that technology, at the bare minimum, should not reinforce oppression through unjust structures and practices of discrimination because it

[1]For an overview of the philosophical debate see Christiano and Bajaj (2024). A central publication in the field of political science is Dahl (2020).

should not violate the principles of democratic equality. It also provides a baseline from which we can evaluate solutions proposed to prevent language technology to participate in discriminatory practices.

2 The Case Study - Methodology

As explained in the introduction, our first goal was to investigate whether and in what sense there is indeed a focus on a techno-fix when it comes to addressing discriminatory behavior of NLP technology. In order to answer these questions we performed a critical interpretative literature review. This type of review integrates the methods used in systematic reviews with a qualitative tradition of inquiry. It includes a transparent methodology for how sources of information are selected, but it does not aim to include all relevant sources Palmer (2022). For our purposes a full coverage of the literature is not necessary, because we are looking for general developments in the methodology applied in NLP research.

The assessment of the literature was performed in two steps. First, we needed to filter out articles that are not on our target topic. Then, we qualitatively assessed the selected articles with respect to our research questions.

Step 1: Filtering For our case study we used the ACL anthologies[2] and focused on all papers published in 2022. This is a choice that can be debated. Choosing ACL is in itself already sampling the publications on language technology. Not all the venues and journals publishing relevant research are included. The question is whether this choice is biased and affects our conclusions. We will come back to this in the discussion section 4. Because of the rapid developments in NLP research, we decided to focus on one year in particular, as this approach may facilitate later studies examining the field's evolution over time.[3] From the 84.000 ACL publications from 2022, we step-wise filtered out all publications that studied specifically the role of NLP systems in discriminatory practices. We filtered all titles and abstracts using a list of keywords. For this we selected a list of words used in the context of describing discrimination (see Table 1). After removing duplicates, a list of 646 papers remained. Two encoders read the abstracts of these papers to determine whether the paper was really addressing unethical behavior of an NLP system. If the annotators disagreed or couldn't decide based on the abstract

[2]https://aclanthology.org/

[3]Our results can be compared to the work of Blodgett et al. (2020) that studies similar questions, but only looking at publications up to 2019.

keywords	"stereotype", "stereotypes", "bias", "fairness", "injustice", "discrimination", "disparity", "disparities", "equality", "equity".

Figure 1: Keywords used for filtering the full body of papers.

whether discrimination were emphasized, they continued reading the introduction, and discussed the article, until an unanimous decision was reached.

The filtering process resulted in a corpus of 207 ACL publications. From this corpus we took a random sample of 50 papers which were then coded with respect to the research questions.[4]

Step 2: Coding The purpose of this case study was to investigate whether there is indeed a focus on a techno-fix in NLP research on unethical behavior of AI. To answer this question we needed to map rather generally how researchers engage with the problem of discriminatory NLP technology. In order to do so we split our main research question into three sub-questions.

1. How do the authors describe the problematic behavior of the technology they are addressing?

2. What are the causes or mechanisms that are given to underlie the problematic behavior of the technology?

3. What solutions are proposed?

The first question targets the framing of the problem. We were interested in seeing whether the authors describe the issue as an ethical issue and what kind of ethical terminology they use to do so. The concern here was that a primary focus on the technological aspects of the problem might lead to a neglect of the ethical and conceptual grounding of the work. The second question aims to map what the authors consider to be possible causes of algorithmic discrimination. These causal dependencies are a good indicator of the solution space they are considering. If algorithmic discrimination is framed as a technological problem, we should find predominantly technological causes mentioned here. The final question targets the solutions that are actually discussed in the papers. A focus on a tech-fix should be clearly visible in the type of solutions here given.

[4]Some of the papers selected did not fit our criteria when we looked at the body of the papers. These papers where replaced by other random choices from the database. In the end, we replace 14 papers from the initial sample for this reason. We will come back to this in the discussion.

We then formulated label groups for each of these sub-questions (for an overview of the labelss see the Appendix). For the first question we defined three label groups: (i) do the authors make a normative/ethical statement, (ii) do the authors provide an explanation for the normative/ethical claim made, and (iii) is the explanation grounded in the theoretical literature. For the second questions, we started out labeling the type of cause for the wrongful behavior of the AI system. Because of the strong focus on bias in an AI system as cause, we also additionally labeled the causes given for this bias. The rationale behind this choice was that this would allow us to trace back the mechanisms taken to be responsible for the problematic behavior of the technology through multiple steps. For the last question, we labeled the type of solution for all solutions discussed in the paper. Additionally, we marked the solution that the authors selected to work on.

We used ATLAS.ti was used for the coding of the articles. For the four main label groups *ethical grounding, cause of wrongful behavior, cause of bias* and *solution* we started with a very fine-grained system of labeling and then merged labels when patterns started to emerge. The full set of labels can be found in the Appendix.

Two encoders read the introduction, the discussion, the conclusions and any other passages of the paper that contained occurrences of the (see Table 1).

3 The case study – Results

3.1 Sub-question 1: conceptual and normative grounding

When looking at how scientists describe the problematic behavior of NLP systems we could distinguish four groups of descriptions (for examples see the appendix):

(i) The system behavior that the paper addressed was never described as negative or unwanted. These were papers that discussed bias measurement/analysis in NLP systems without stating that bias could potentially be problematic.

(ii) The system behavior the paper addressed was described as negative, but not explicitly as ethically wrong. In this case either (a) terms such as "unwanted" or "negative" were used without clarifying that the behavior was morally problematic, or (b) the discussion focused on legal violations rather than ethical considerations,[5] or (c) the authors referred to issues commonly

[5]We gathered reference to law violations under this label, because it concerns adherence to a different type of norm: you can avoid doing something because it is against the law, without that you have to believe that it is morally wrong. Of the three papers with that label two concern privacy issues. In both cases it was not entirely clear whether privacy was clearly seen as an ethical or only a legal issue.

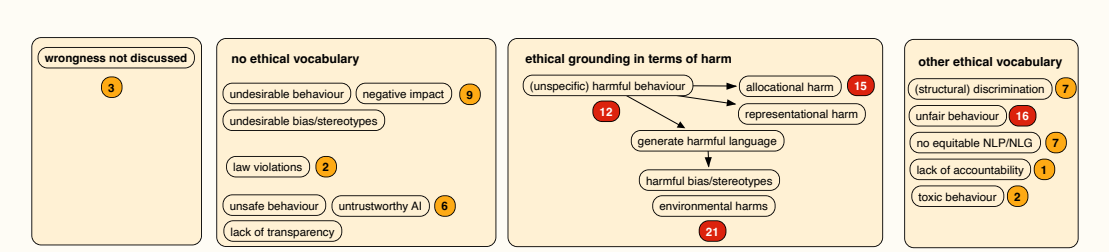

Figure 2: The figure displays the labels used to code ethical grounding. From left to right the ethical vocabulary becomes more developed. The numbers next to a label group reports the number of papers in which this label/label group is used. The red numbers represent the dominant labels/label groups in the sample. The arrow represents label-dependence. A single paper could be labeled with different quote-level labels.

associated with unethical AI — such as unsafe systems, lack of trustworthiness, or insufficient transparency — without framing these concerns explicitly in terms of moral wrongdoing.

(iii) The negative impact of technology was described in terms of harm.

(iv) Other ethical vocabulary was used to describe the negative impact.

An overview of the labels in each group, together with the counts of papers in which the label occurred is given in Figure 2. There is a strong dominance of framing the problematic behavior in terms of harm. The harm concerned nearly exclusively harm caused by the output of the system. A single paper in the sample referred to harm that can occur during production of AI systems ([Sample13]). Apart from harm, also framing in terms of fairness was frequent. Other related vocabulary is also used but much less frequent. Furthermore, we noticed that there is still as substantial group of papers that do not frame the problem as ethically wrong, but instead as undesired or make no clear normative statements.

We also often see that a paper makes a normative claim, but without any additional explanation about why the behavior is wrong, to whom and in what situation. An example is [Sample4] , a paper that discusses measures of social biases in prompt-based Multi-Task learning. The only normative statement in the Introduction section occurs in the description of the contribution of the paper: "BBNLI measures human cognitive biases across three different domains and capturing 16 *harmful stereotypes*." (emphasis added by the authors). Nowhere in the paper is it explained what biases are, how they are related to stereotypes, and what is wrong about biases and stereotyping in general. From the fifty papers in

the sample, only 22 papers attempt to explain at least once the normative/ethical claims made.

There is also limited grounding in interdisciplinary literature on discrimination. Only twenty-five of the fifty papers referred to non-technologial work on the topic.[6] Seven of these papers quoted Blodgett et al. (2020) in this context. We noticed that grounding occurred more frequently in case the wrong was framed in terms of fairness. This is probably the case because there exists a broader spectrum of accessible literature on the issue of fairness than for bias and harm. Also papers using the terms "allocational harm" or "representational harm" tended to be more explicit about their normative grounding. In this case a very clear reference exists (Barocas et al. (2023)) that can provide guidance.

Even in case a paper did offer some explanation for how or why the targeted behavior of an AI system was wrong, this explanation was in general not very well developed and the literature provided very limited. For instance, we labelled the quote "... language models (LMs) that empower much of the web as we know it are well known to contain biases that promote structural discrimination in downstream tasks against minorities and larger social groups alike (Bender et al., 2021)." [Sample23] as providing a normative explanation, because a link is made between biases and discrimination in downstream tasks. However, the paper does not define the concept of bias or explain what is meant by structural discrimination in downstream tasks, linking to the substantial philosophical and socio-theoretical work on the matter. Thus, the key concepts of the research are still poorly grounded.

3.2 Sub-question 2: Causes of wrongful behavior

In the next step, we looked at what causes were given in the papers for the unethical or unwanted behavior of language technology. We could distinguish four groups of root causes that researchers pointed to: (1) the training of the technology, (2) technological causes, (3) causes related to the research or development agenda and (4) causes related to the use of the technology (see the bold labels in Figure 3; examples for all label groups can be found in the appendix). The cause stated most often offered was bias, either bias in the training data (issues with training) or biased behavior of the AI technology (technological causes). If researchers pointed to issues with the training data as the cause of the wrong behavior of the AI system, the wrong was described as bias in the system. As the reader can see, very few papers mentioned other causes. In most cases bias was stated as the

[6]There were papers that did not provide an explanation for their normative statements, but did cite theoretical literature. For instance, see [Sample9] : "Deploying such biased word embeddings in downstream tasks would cause allocational and representational harms (Blodgett et al., 2020).".

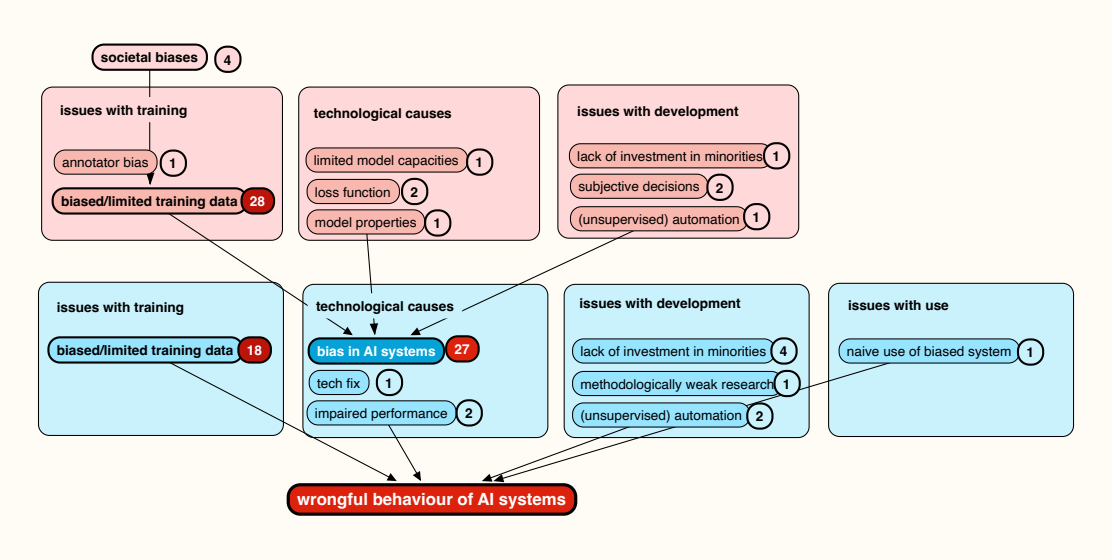

Figure 3: The figure displays the labels used to code the causes discussed in the papers for the wrongful behavior of NLP technology. The upper, read labels code causes provided for bias in AI systems, if this was listed as cause of the wrongful behavior. The numbers next to a label group give the number of papers in which this label is used. The red numbers represent the dominant labels in the sample. Arrows represent causal dependence

cause, without that it was made clear what exactly was meant by bias and how this bias caused the unethical behavior.

We also labeled causes of bias in AI systems if the latter was mentioned as cause of the wrongful behavior, see the red boxes in Figure 3. Here, most authors saw biased or limited training data as cause for the bias in the AI system, see, for instance [Sample40] : "Disparities may reflect data imbalances: If training data contains less data from a group, predictions for that group will tend to be worse.", or [Sample26] : "... there is a growing concern if such language models contain social biases such as stereotyping negative generalizations of different social groups and communities, which might have been present in their training corpora Liang et al. (2021); Garrido-Muñoz et al. (2021).". Very few alternative causes were discussed. The same classification emerged as with the primary cause of the unethical behavior: the cause was either located in the training, or technological issues were blamed, or the research or development agenda. Issues with use were not relevant in case we look at bias in an AI system.

In sum, the causal narrative that emerges is relatively straightforward and predominantly centered on the technology itself: the wrongdoing is attributed to a biased AI system, which in turn is biased due to limitations/bias in the training

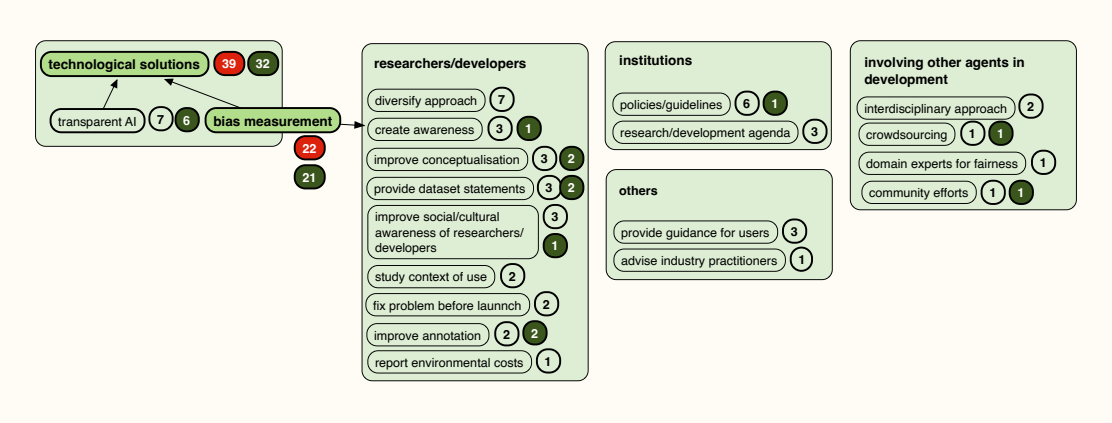

Figure 4: The figure displays the labels used to code solutions. The numbers next to a label report first the number of papers in which this solution is mentioned and, secondly, the number of papers in which this is the solution the paper works on. Papers can work on multiple solutions (for instance, measuring bias and mitigating bias). The red numbers represent the dominant labels.

data. This framing subsequently shapes the space within which researchers seek solutions to the system's problematic behavior.

3.3 Sub-question 3: Solutions proposed

Lastly, we labeled all solutions mentioned by the authors of the articles. We, again, found four groups of solution types: 1) purely technological solutions, 2) solutions that addressed the behavior of the researcher/engineer, 3) solutions concerning institutional behavior, and 4) solution targeting the behavior of other agents. We found that technical solutions strongly dominated, as can been seen in Figure 4 (again, examples for all label groups can be found in the Appendix). Additionally, nearly all the papers focused in their research only on technical solutions. We labeled papers that worked on bias measurement separately, even though these are also technological solutions, because bias measurement can also be pursued to raise more general awareness of the problem. However, in most cases, even if awareness was given as reason to develop bias measures, this was described as awareness of the researchers, and not those of the user or other agents. For example, in [Sample3] the focus is on awareness, but this is awareness of design choices when measuring bias in language generation. [7]

There was only one paper in the sample that did not present a technological

[7] [Sample24] appears to be an exception and targets general awareness of the limitations of language models.

solution/bias measurement as its main contribution (i.e. [Sample40] , in which regulative measures for the employment and development of NLP technology are discussed[8]). Furthermore, the focus of nearly all solutions is on fixing the system. Improving annotation processes or involving other agents in development is in the end still for the purpose of building technology that solves the issue of algorithmic discrimination. This applies to nearly all the solutions proposed under the label groups *researchers/development*, *institutions*, and *involving other agents in development*.

3.4 Conclusion

With respect to our first question we see from our results that the conceptual grounding of the normative claims made in the literature is still weak. While more than 2/3 of papers use ethical terms to describe the problematic behavior of the NLP technology they discuss, the description remains rather superficial and not well anchored the in theoretical literature. Concretely, in the majority of the papers nothing is said about why the system's behavior is harmful, to whom and in what context etc..

With respect to our second question concerning the causes discussed for the wrongful behavior of the AI systems we observe a rather limited and uniform view on the causes. Furthermore, there is a clear focus on locating the problem in the technology and blaming bias for it. We however found very little reflection on the interaction between technology and society or the mechanisms that determine developmental agendas or choices of datasets.

The solutions discussed strongly focused on technical solutions. Most of them concerned bias measurements. Other solutions were considered, but rarely the focus of the work reported on. Many of those solutions also served the purpose of improving the technology.

In sum, our finding indicate that, in 2022, NLP-research on discrimination was strongly oriented toward technological solutions. The research landscape seems to form a techno-centric research bubble with a very clear consensus on what a relevant research questions are and how they should be addressed. The predominant focus was on identifying and mitigating bias in NLP systems using technological approaches. There was limited concern for the complex interaction between technology and society and the majority of studies lacked substantial theoretical grounding in socio-technical or ethical frameworks.

[8]This also the single paper discussing discrimination through the production of NLP technology.

4 Discussion

4.1 Methodological limitations

The validity of our results depends on the representativeness of our sample. There are large conferences addressing ethical and societal implications of AI technology like FAccT, AIES and EAAMO that are not included in the ACL anthology. However, these conferences are not specifically about NLP research, which is the topic of the present manuscript. This is why we opted for using ACL. But there will be papers addressing language technology that went to these conferences specifically because they focused on societal and ethical issues. In future work we will extend the search space for relevant papers.

There are two steps in the selection of the sample from our database that could have biased the sample. First of all, we might have chosen the wrong set of keywords for the first filtering. The worry here would be that our set misses key terminology used in the debate of discrimination in NLP and, therefore, we filtered out part of the discourse. However, these terms would have come up during close reading of the texts, which did in fact happen for "harm". This was the only additional term that we found during close reading of the sample. We therefore think that the risk that we missed an important part of the discourse that would change our results is small.

The second error-sensitive step is the manual filtering and annotation of the sample. We already noticed in footnote 2 that we had a high rate of false positives in the set of papers selected as addressing discrimination. Although false positives were filtered out in later stages of the selection process and thus are unlikely to have significantly affected the results, their high number raises concerns about the potential presence of false negatives. However, we observed that most of the false positives, i.e. papers initially classified as relevant, mention the societal impact of the discussed technology in the abstract, but do not return to this point in the full text. This observation supports the assumption that if a paper genuinely engages with issues related to algorithmic discrimination, such relevance would likely be reflected in the abstract, given that societal impact is often used as a selling point. Therefore, we consider the likelihood of false negatives to be low.

4.2 Putting the results in perspective

If we agree that there is a strong focus on a techno fix in the research literature, we still do not need to agree that this is problematic. One might argue, for instance, that finding a focus on technological solutions is neither surprising, nor problematic given that we looked at a sample of papers written by NLP engineers and researchers. So, of course they will focus on technological solutions.

Let us draw a distinction between the group of NLP practitioners and the field of NLP. NLP practitioners, as we understand them here, are the people who work on the technological part of NLP systems. These are the code developers, the data scientists and engineers that do the actual implementation and training of the systems. Apart from the practitioners, however, we have the *field* of NLP, which is the study of NLP technology, naturally in relation to the society that builds and employs this technology. The standards for what is done in the field of NLP should not be set by what NLP practitioners could do, but by what needs to be done to tackle the problems we face. Developing technological fixes for algorithmic discrimination is not the only, and arguably also not the best way to address the issue. Consequently, we should find work on such alternative solution in the field. Our case study suggests that this work is missing.

A lot needs to be invested into NLP research to make sure that we develop technology that is beneficial to society. But this involves more than just engineering. We are certainly not the first to point this out. In Blodgett et al. (2020), for instance, very concrete proposals for research questions are made that need to be answered. However, NLP practitioners cannot answer these questions by themselves.[9] This needs to be an interdisciplinary effort, in dialogue with society.

This also answers a second potential objection: Why not adopt a techno-centric approach to algorithmic discrimination as long as it solves the problem? The issue with this objection is that despite an immense investment of research in recent years, so far we do not have technical solutions. One possible explanation for the lack of progress is a lack of understanding of the problem. We would argue that our understanding of the problem is deficient exactly because we try to understand it as a purely technological challenge. Appreciating algorithmic discrimination as the social-technological problem it is will open new perspectives on how it can be addressed.

This is where we hope to make a contribution in the final part of this manuscript. We believe that the focus on a techno-fix is at least partly due to the particular way in which the problem of algorithmic discrimination is currently framed in the field of NLP. In the next section we will rethink this framing and propose an alternative language to ground work on algorithmic discrimination.

[9] Just to give an example, the first of the questions proposed in Blodgett et al. (2020) is: "How do communities become aware of NLP systems? Do they resist them, and if so, how?". This is a question that requires research methodology that is not in the toolbox of an NLP practitioner.

5 How to move forward

5.1 The delusive lure of the techno-fix

The observations we made in our case study are not new. Very similar results have been reported, for example, in Blodgett et al. (2020), a study published in 2020. Over the last couple of years various authors have expressed criticism about the focus on quick technological solutions. One might wonder why there is so little change in this direction, while at the same time the number of publications on algorithmic discrimination grew immensely.

There are various reasons for this. Here we want to focus on two misguided assumptions that make the techno-fix so appealing. The first misguided assumption is well known. It is the assumption that technology is inherently value-free, that not only can complex social issues be reduced to a technological one, but that this can be done because of technology's objectivity and capacity for impartial decisions. This conviction ignores the very real ways in which our tools and technologies reflect and embody the environment they emerge from. These things do not spring from nowhere. Every step of the creation process, from how data is collected to how a tool is implemented and used in the world, contains and expresses underlying societal arrangements. This is not specific to AI technology. Winner (1980) showed how certain infrastructure and architectural designs inherently include some and exclude others, thereby embodying pre-existing political dynamics and reinforcing inequality through their own effects. Benjamin Benjamin (2019) makes this point in connection with the topic of this paper by thoroughly describing how racial constructs become materially embodied in technologies, and then contribute to discrimination of minorities.

Technology is always situated. It emerges from certain contexts and has certain effects, whether intended or not. As such, any technological assessment cannot be simplified to looking at the tool in isolation. Instead, it must be understood as an ecosystem Stahl (2022, 2023), as an active element within a larger environment of values and power relations. This is crucial not only for any comprehensive ethical evaluation, but – again – for the very functioning of democracy.

We want to add a second misguided assumption here, which is not yet part of the discussion: algorithmic discrimination is generally treated as a *safety issue*. There is an important difference between developing technology that does not contribute to discrimination and, say, integrating safety checks into an automatic door. We can easily agree on what it means for such a door to operate safely. However, as society, we are still struggling with finding a ground-truth when it comes to discrimination. Relying on technocratic solutions or Big Tech to resolve problems like algorithmic discrimination sidesteps valuable democratic processes and hinders societal growth Weinberg (1956); Birhane (2021). How to achieve

respectful, fair and just technologies cannot and should not be decided for us, we, i.e. society, should decide the meaning of these concepts ourselves. After all, once these meanings are formalized into technology, they become reproduced in society at an incredibly large scale.

Both assumptions lead us to believe that solving algorithmic discrimination is just a matter of fixing the technical system, and that these solutions will benefit society at large. Instead, we need to realize that technology can only be developed in dialogue with society. It needs to be grounded in the academic and societal debate on norms and values and the ethical/political assumptions build into technology need to be transparently presented.

This is directly related to one of our findings in the case study: the normative/ethical grounding of the research done is still weak. In the remainder of this section we therefore want to provide some tools that can help with the ethical and normative grounding of research on algorithmic discrimination by looking at the vocabulary presently in use and the implicit value-sensitive choices that have already been made here.

5.2 The limitations of harm and bias

In our case study, we saw that the terms *bias* and *harm* heavily dominated the conceptual discussion in the articles we reviewed (see Figure 2). However, the usefulness of these terms in providing solid normative anchors for work on algorithmic discrimination is rather limited.

Let us first look at *bias*. Although this term is often used with a normative connotation (bias is bad), it is not an inherently ethical notion. Finding a bias in some system's behavior is not necessarily related to an actual wrongdoing. The ethical grounding still needs to be elaborated on; questions like what is wrong about this bias and why, who is wronged, and how; still need to be answered. Thus, while *bias* is an important and useful notion in the debate, it does not provide a normative grounding by itself.

Harm fairs somewhat better. Using harm allows for an intuitive and immediate grounding of any norms one might want to propose – of course, technology should not do harm; the user should not suffer in any sense. Still, there are also issues with this notion. Some of these problems have been already pointed out by others. For instance, one still needs to define what exactly is meant by harm, who is harmed, and in what situations. In order to understand the problem, make a normative assessment, and develop solutions, this is all important information; just pointing to the potential of harm is not sufficient. In addition, the system behavior under scrutiny needs to be thoroughly related to harm. For example, take the very common claim in our sample that harm is caused by bias in language models. This is often all that is said about what is problematic about model bias. The authors

proceed by developing and/or evaluating some measure for this bias. However, pointing to performance on these measures is not equivalent to showing that harm will or will not occur, because the relation between these abstract measures and actual harm is at best unclear.

There are some more fundamental issues with harm that should be considered. Most importantly, the concept of harm may not offer as firm an ethical grounding as initially imagined and, therefore, cannot be used to fully explain what is wrong about discrimination. There is a conceptual difference between harm and wrong Feinberg (1987). Only the latter possesses moral weight since not all harms are necessarily wrongs. Imagine someone punished for disobeying the law or even someone who is harmed by a coconut falling on their head. These are cases in which an individual is harmed, however whether they have been wronged is not as clear, and this distinction has serious repercussions on questions of moral retributions and responsibilities. A harm becomes a wrong when an action is unjustified, only then does it become morally indefensible, and calls into question the relationships between agents involved and what they owe each other Diberardino et al. (2024). One consequence of this is that harm is not sufficient for discrimination to take place.

But it is not just that the notion of harm is not sufficient for discrimination, it also is not necessary. Discrimination can occur without there being a clear sense of harm done. Take the case of the Kenyan workers who get paid 2$ per hour to clean toxic content from the output of ChatGPT Perrigo (2023). In the local context of Angola this is a well-paid job. Therefore, there is certainly no easy way to argue for economic harm in this case. One could argue that there is psychological harm, depending on the content the workers need to clean up Block and Riesewieck (2018). But even in the case that there is no psychological damage involved in this kind of work, it is still morally problematic because it is unjust. These are practices that grew out of post-colonial structures that exploited human beings and lead to an unfair distribution of opportunities and goods of which the effects still resonate today Hammer and Park (2021).

Only taking psychological harm into account may also deflect attention from the true source of the injustice, that is, interlocking systems of oppression and the unjust social practices which support them. One response might be to remedy the psychological damage with individual band-aids such as 'free' employee therapy sessions or a complementary yoga retreat from Meta, OpenAI or one of their subcontractors. But this is not the right angle to address the problem. What is causing the harm in this case is not what is causing the injustice.

This example illustrated that framing negative social impact in terms of harm does not provide a proper viewpoint for adequately addressing the wrongs of these technologies and the unjust practices in which they are embedded. The focus on

harm (similar for fairness) is connected to a focus on system-user interaction as the locus of trouble: discrimination takes place from output of the system to the agent using it. This is an individualistic framing of the problem of discrimination, which we also see outside the AI-debate. Discrimination is predominantly framed as interpersonal: one person or agent discriminating against another. The cause of discrimination then is located in the individual beliefs and attitude of a person versus another Becker (1971); Garcia (1996); Correll et al. (2010). Similarly, the problem of algorithmic discrimination is limited to the NLP system that incidentally might wrong a user or group of users because of prejudiced 'beliefs' encoded in its internal states. The similarities even extend to blaming the training data (childhood, education) for the problematic 'beliefs' (see Figure 3 and the dominance of bias in the data set as cause of system bias). The solution then often lies in merely correcting these biases, again something that clearly showed up in our data (see Figure 4).

To sum up this part of the discussion, harm is neither necessary nor sufficient for discrimination. Most importantly, the focus on harm leads to an interpersonal understanding of discrimination, and reducing discrimination to the interpersonal level creates blind spots for the structural dimensions of discrimination. We are not arguing to get rid of the notion of harm altogether. On the contrary, there is an important role for harm to play in the context of the debate on algorithmic discrimination. It is one of the central reasons for what is wrong about discrimination. It is just not the only one. In other words, we should keep the notion of harm as referring to damage or suffering that might be caused to particular groups or individuals, but we need a different, more general and morally consequential notion to ground the discussion of algorithmic discrimination.[10]

5.3 Injustice instead of harm

Let us go back to our beginning. The problem we are trying to solve here is algorithmic discrimination. Discrimination is treating groups or individuals differently in a way that is morally wrong. In the Introduction we linked this moral wrongdoing to violations of the foundations of democracy.

It is not making a distinction that is problematic – we have to treat different people differently. It is also not that we harm one group and not the other that is

[10]One possible candidate for such a notion is fairness. We noticed, for instance, that approaches in our sample that focused on fairness achieved a more thorough conceptual grounding of their normative claims, because of the rich literature that exists on formalizations of fairness Narayanan (2018) However, we also observed that articles nearly exclusively focused on these formalizations without considering the actual social processes involved (see also Taylor (2023)). We see this as a weakness in the fairness approach and believe it can be fortified with the framing we put forward here.

at issue – such a distinction can be justified. What is wrong about discrimination is that it is unjust. This latter notion, injustice, lays at the heart of algorithmic discrimination.

Therefore, we propose to replace the notion of harm with that of injustice. Of course, there are also debates regarding definitions of in/justice, and formalizations may be just as varied and complex as those trying to formalize a notion of bias Edenburg and Wood (2023). However, tackling the complexity of injustice instead of bias would urge researchers using the notion to dig deeper when providing the conceptual and normative grounding of their research rather than resorting to harm or discrimination full stop.

Here, we want to propose to use work on injustice by Haslanger Haslanger (2022, 2023, 2024), which naturally incorporates many aspects that came up in our discussion. A just system then, according to Haslanger would be one in which the practices and structures of this system would enable individuals to reflect critically on what they value and to pursue these values. This means that the system should support and ensure the social and material conditions for the development of skills, knowledge and resources that this requires. In other words, a just system supports people's full democratic participation Haslanger (2024). A concern for justice is what grants equal opportunities to groups and individuals, it is what guides true democratic participation and deliberation, and it must be practiced at every step of technological development if we want the scientific community to be an ethical and politically responsible contributor to democratic values.

Haslanger's framework also naturally incorporates the structural and systemic aspects of discrimination that were missing when focusing on the notion of harm. She states that 'structural injustice occurs when the practices that create the structure – the network of positions and relations – (a) distort our understanding of what is valuable or (b) organize us in ways that are unjust/harmful/wrong, e.g., by distributing resources unjustly or violating the principles of democratic equality' Haslanger (2022). Following this, systemic injustice 'occurs when an unjust structure is maintained in a complex system that its self-reinforcing, adaptive, and creates subjects whose identity is shaped to conform to it' Haslanger (2022).

To recognize that structural injustice is a fundamental aspect of discrimination subsequently requires more than removing bias from a model or dataset. Even a technology that operates accurately and does not show biased behavior in its user interactions can still contribute to practices that produce unjust social relations. Let us take facial recognition technologies, which often show biases and discriminate against people of color, women, and, women of color in particular Benjamin (2019); Noble (2018); Buolamwini and Gebru (2018); Perkowitz (2021); Hill (2022). These biases have been linked to biased or limited training data of these systems. This, in turn has led to calls for more thorough training of these

systems on a greater variety of data. However accurate facial recognition can still have discriminatory effects because it can reinforce an unjust structure as part of an oppressive system. This system organizes and (re-)produces subjects based on certain markers, like gender, race, species, ability and so on, which are politically loaded categories. Even when we succeed in debiasing the technology, the technology is not off the moral hook. Making the dataset a more accurate and balanced representation of people might even increase harm for a minority producing an effective tool to enforce (and legitimize!) an oppressive system Benjamin (2019). While one might argue that this merely points at a problematic use of the technology being targeted at an oppressed minority, this foregoes the fact that these systems are inherently unjust because they rely on categorization based on facial features including skin color.

To summarize, harm does not give us sufficient ethical or explanatory power. When a harm is also a wrong, the normative dimension becomes irrefutable. Wrongs are inherently unjustified and demand reflection on our moral obligations and responsibilities to each other. Yet injustice as we have described it so far, points at a distinct normative wrong from the perspective of democratic participation. It lets us position our considerations in concrete democratic values and practices. At the beginning we have mentioned that we ground our normative position with respect to discrimination in adopting democracy as a core value. This means embracing equal rights and equal opportunities for all members of society. Since systemic and structural injustice violate these ideals of equality and democracy, using a concept of injustice when framing the problem of algorithmic discrimination gives us both the space and strength needed to make substantive ethical and political arguments for assessing technology.

6 Conclusions

In this article we have been taking a meta-perspective, studying practices within the field of NLP in 2022. The goal was to engage with the critique that the field overly focuses on technological solutions when addressing the problem of algorithmic discrimination. Our case study has confirmed this suspicion. It also points to a possible explanation for this limitation: a too narrow framing of the problem, together with a shallow grounding of its ethical and societal dimensions. In the second part of the paper we wanted to provide some tools that could help with finding a more substantial ethical understanding of what algorithmic discrimination entails. Concretely, we have proposed to replace the focus on harm in the debate with one on injustice. Using one particular way to flesh out what injustice means, building on the work of Haslanger, we showed that this allows a broader perspective on the underlying dynamics leading to algorithmic discrimination. An

important part of this is understanding the role technology can play in structural, and even systemic forms of discrimination. This framework also highlight the close relationship between discrimination and impeding core practices of a democratic society.

Looking at the problem of algorithmic discrimination from a systemic point of view also implies seeing the NLP practitioners themselves as embedded in social structures. This can mean paying attention to the well-discussed fact that NLP practitioners have certain demographic features and that this affects the way they set research agendas and build AI systems. This is, however, not the only way structural factors influence their work. NLP researchers are embedded in various power structures that steer their behavior. They face significant pressures: to publish, gain conference acceptance, and secure funding. For those working in industry these demands are even more adamant by an additional layer of institutional and commercial constraints. While these pressures shape the current landscape of NLP research, they should not be perceived as immutable limitations. Instead, we must recognize these power structures and use them as a starting point for a discussion on how they can be challenged. In the end, it is the research community itself that defines what counts as good science – and that means it also holds the power to redefine it.

References

A. Balayn and S. Gurses. 2021. *Beyond Debiasing. Regulating AI and its Inequalities.* Technical Report. European Digital Rights (EDRi), Brussel, Belgium.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities.* MIT Press.

Gary S. Becker. 1971. *The Economics of Discrimination.* The University of Chicago Press.

Ruha Benjamin. 2019. *Race After Technology. Abolitionist Tools for the New Jim Code.* Polity Press.

Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns* 2, 2 (2021), 100205. `https://doi.org/10.1016/j.patter.2021.100205`

Hans Block and Moritz Riesewieck. 2018. The Cleaners. GEBRÜDER BEETZ FILMPRODUKTION KÖLN GMBH & CO. KG. and Filmart Produções.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceed-*

ings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification.. In *Conference on fairness, accountability and transparency*. PMLR, 77 – 91.

Tom Christiano and Sameer Bajaj. 2024. Democracy. In *The Stanford Encyclopedia of Philosophy* (Summer 2024 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University.

Joshua Correll, Charles M. Judd, Bernadette Park, and Bernd Wittenbrink. 2010. Measuring prejudice, stereotypes and discrimination. In *The SAGE handbook of prejudice, stereotyping and discrimination.* Sage Publications Inc, 45 – 62.

Robert A Dahl. 2020. *On democracy*. Yale university press.

Nathalie Diberardino, Clair Baleshta, and Luke Stark. 2024. Algorithmic Harms and Algorithmic Wrongs.. In *FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil). Association for Computing Machinery, New York, NY, USA.

Elizabeth Edenburg and Alexandra Wood. 2023. Disambiguating algorithmic bias: from neutrality to justice.. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics and Society*. 691 – 704.

Joel Feinberg. 1987. *The Moral Limits of the Criminal Law Volume 1: Harm to Others.* Oxford University Press.

Jorge J. E. Garcia. 1996. The Hart of Racism. *Journal of Social Philosophy* 27 (1996), 5 – 45. Issue 1.

Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L Alfonso Ureña-López. 2021. A survey on bias in deep NLP. *Applied Sciences* 11, 7 (2021), 3184.

Ricarda Hammer and Tina M. Park. 2021. The Ghost in the Algorithm: Racial Colonial Capitalism and the Digital Age. *Global Historical Sociology of Race and Racism Political Power and Social Theory* 38 (2021), 221 – 249.

Sally Haslanger. 2022. Systemic and Structural Injustice: Is There a Difference? *Journal of Social Philosophy* 53 (2022), 512 – 534. Issue 4. https://doi.org/10.1111/josp.12373

Sally Haslanger. 2023. Systemic and Structural Injustice: Is There a Difference? *Philosophy* 98 (2023), 1 – 27. Issue 1. https://doi.org/10.1017/s0031819122000353

Sally Haslanger. 2024. Social Structures as a Site for Injustice: Why Social Theory Matters. (November 2024). Talk given at the 2nd Workshop on Algoritmic Injustice (WAI2).

Kashmir Hill. 2022. Wrongfully Accused bu an Algorithm. In *Ethis of Data and Analytics.* Auerbach Publications.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*. PMLR, 6565–6576.

Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics.. In *Conference on Fairness, Accountability and Transparency* (New York). 3.

Safiya Noble. 2018. *Algorithms of Oppression: How Search Engines Reinfornce Racism.* New York University Press.

M. A. Palmer. 2022. Qualitative Synthesis Methods: Critical Interpretive Reviews, Narrative Reviews, Expert Opinions. https://www.sesync.org/resources/qualitative-synthesis-methods-critical-interpretive-reviews-narrative-reviews-expert. Accessed: 2024-04-10.

Sidney Perkowitz. 2021. The Bias in the Machine: Facial Recognition Technology and Racial Disparities. *MIT Case Studies in Social and Ethical Responsibilities of Computing* Winter 2021 (feb 5 2021). https://mit-serc.pubpub.org/pub/bias-in-machine.

Billy Perrigo. 2023. Exclusive: OpenAI Used Kenuan WOrkers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. *Time Magazine* 18 (2023).

John Rawls. 2001. *Justice as fariness: a restatement.* The Belknap Press of Harvard University Press.

Bernd Carsten Stahl. 2022. From computer ethics and the ethics of AI towards an ethics of digital ecosystems. *AI Ethics* 2 (2022). https://doi.org/10.1007/s43681-021-00080-1.

Bernd Carsten Stahl. 2023. Embedding responsibility in intelligent systems: from AI ethics to responsible AI ecosystems. *Scientific Reports* 13 (2023). Issue 1.

Linnet Taylor. 2023. Can AI governance be progressive? Group interest, group privacy and abnormal justice. In *Handbook on the Politics and Governance of Big Data and Artificial Intelligence.* Edward Elgar Publishing, 19 – 40.

Alvin M. Weinberg. 1956. Basic Research and National Goals: A Report to the Committee on Science and Astronautics, U.S. House of Representatives, by the National Academy of Sciences. *Minerva* 3, 4 (1956), 499–523.

Langdon Winner. 1980. Do Artefacts Have Politics? *Daedalus* 109 (1980), 121–136. Issue 1.

Iris M. Young. 1990. *Justice and the Politics of Difference.* Princeton Univesity Press.

Appendix

Articles in the study sample

[Sample1] Arshiya Aggarwal, Jiao Sun, and Nanyun Peng. 2022. Towards Robust NLG Bias Evaluation with Syntactically-diverse Prompts. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 6022–6032. `https://doi.org/10.18653/v1/2022.findings-emnlp.445`

[Sample2] Jaimeen Ahn, Hwaran Lee, Jinhwa Kim, and Alice Oh. 2022. Why Knowledge Distillation Amplifies Gender Bias and How to Mitigate from the Perspective of DistilBERT. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, Christian Hardmeier, Christine Basta, Marta R. Costa-jussà, Gabriel Stanovsky, and Hila Gonen (Eds.). Association for Computational Linguistics, Seattle, Washington, 266–272. `https://doi.org/10.18653/v1/2022.gebnlp-1.27`

[Sample3] Afra Feyza Akyürek, Muhammed Yusuf Kocyigit, Sejin Paik, and Derry Tanti Wijaya. 2022a. Challenges in Measuring Bias via Open-Ended Language Generation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, Christian Hardmeier, Christine Basta, Marta R. Costa-jussà, Gabriel Stanovsky, and Hila Gonen (Eds.). Association for Computational Linguistics, Seattle, Washington, 76–76. `https://doi.org/10.18653/v1/2022.gebnlp-1.9`

[Sample4] Afra Feyza Akyürek, Sejin Paik, Muhammed Kocyigit, Seda Akbiyik, Serife Leman Runyun, and Derry Wijaya. 2022b. On Measuring Social Biases in Prompt-Based Multi-Task Learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 551–564. https://doi.org/10.18653/v1/2022.findings-naacl.42

[Sample5] Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2022. The Arabic Parallel Gender Corpus 2.0: Extensions and Analyses. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 1870–1884. https://aclanthology.org/2022.lrec-1.199

[Sample6] Sarah Alnegheimish, Alicia Guo, and Yi Sun. 2022. Using Natural Sentence Prompts for Understanding Biases in Language Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 2824–2830. https://doi.org/10.18653/v1/2022.naacl-main.203

[Sample7] Saied Alshahrani, Esma Wali, and Jeanna Matthews. 2022a. Learning From Arabic Corpora But Not Always From Arabic Speakers: A Case Study of the Arabic Wikipedia Editions. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, Houda Bouamor, Hend Al-Khalifa, Kareem Darwish, Owen Rambow, Fethi Bougares, Ahmed Abdelali, Nadi Tomeh, Salam Khalifa, and Wajdi Zaghouani (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 361–371. https://doi.org/10.18653/v1/2022.wanlp-1.34

[Sample8] Saied Alshahrani, Esma Wali, Abdullah R Alshamsan, Yan Chen, and Jeanna Matthews. 2022b. Roadblocks in Gender Bias Measurement for Diachronic Corpora. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, and Lars Borin (Eds.). Association for Computational Linguistics, Dublin, Ireland, 140–148. https://doi.org/10.18653/v1/2022.lchange-1.15

[Sample9] Haozhe An, Xiaojiang Liu, and Donald Zhang. 2022. Learning Bias-reduced Word Embeddings Using Dictionary Definitions. In *Findings of the Association for Computational Linguistics: ACL 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 1139–1152. https://doi.org/10.18653/v1/2022.findings-acl.90

[Sample10] Giuseppe Attanasio, Debora Nozza, Eliana Pastor, and Dirk Hovy. 2022. Benchmarking Post-Hoc Interpretability Approaches for Transformer-based Misogyny Detection. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, Tatiana Shavrina, Vladislav Mikhailov, Valentin Malykh, Ekaterina Artemova, Oleg Serikov, and Vitaly Protasov (Eds.). Association for Computational Linguistics, Dublin, Ireland, 100–112. https://doi.org/10.18653/v1/2022.nlppower-1.11

[Sample11] Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. 2022. How well can Text-to-Image Generative Models understand Ethical Natural Language Interventions?. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 1358–1370. https://doi.org/10.18653/v1/2022.emnlp-main.88

[Sample12] Lisa Bauer, Karthik Gopalakrishnan, Spandana Gella, Yang Liu, Mohit Bansal, and Dilek Hakkani-Tur. 2022. Analyzing the Limits of Self-Supervision in Handling Bias in Language. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 7372–7386. https://doi.org/10.18653/v1/2022.findings-emnlp.545

[Sample13] Conrad Borchers, Dalia Gala, Benjamin Gilburt, Eduard Oravkin, Wilfried Bounsi, Yuki M Asano, and Hannah Kirk. 2022. Looking for a Handsome Carpenter! Debiasing GPT-3 Job Advertisements. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, Christian Hardmeier, Christine Basta, Marta R. Costa-jussà, Gabriel Stanovsky, and Hila Gonen (Eds.). Association for Computational Linguistics, Seattle, Washington, 212–224. https://doi.org/10.18653/v1/2022.gebnlp-1.22

[Sample14] Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 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), Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 561–570. https://doi.org/10.18653/v1/2022.acl-short.62

[Sample15] Yang Trista Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 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*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 1276–1295. https://doi.org/10.18653/v1/2022.naacl-main.92

[Sample16] Mayukh Das and Wolf Tilo Balke. 2022. Quantifying Bias from Decoding Techniques in Natural Language Generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 1311–1323. https://aclanthology.org/2022.coling-1.112

[Sample17] Awantee Deshpande, Dana Ruiter, Marius Mosbach, and Dietrich Klakow. 2022. StereoKG: Data-Driven Knowledge Graph Construction For Cultural Knowledge and Stereotypes. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, Kanika Narang, Aida Mostafazadeh Davani, Lambert Mathias, Bertie Vidgen, and Zeerak Talat (Eds.). Association for Computational Linguistics, Seattle, Washington (Hybrid), 67–78. https://doi.org/10.18653/v1/2022.woah-1.7

[Sample18] Fatma Elsafoury, Steve R. Wilson, Stamos Katsigiannis, and Naeem Ramzan. 2022. SOS: Systematic Offensive Stereotyping Bias in Word Embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 1263–1274. https://aclanthology.org/2022.coling-1.108

[Sample19] Yacine Gaci, Boualem Benatallah, Fabio Casati, and Khalid Benabdeslem. 2022. Debiasing Pretrained Text Encoders by Paying Attention to Paying Attention. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 9582–9602. https://doi.org/10.18653/v1/2022.emnlp-main.651

[Sample20] Divyam Goel and Raksha Sharma. 2022. Leveraging Dependency Grammar for Fine-Grained Offensive Language Detection using Graph Convolutional Networks. In *Proceedings of the Tenth International Workshop on Natural Language Processing for Social Media*, Lun-Wei Ku, Cheng-Te Li, Yu-Che Tsai, and Wei-Yao Wang (Eds.). Association for Computational Linguistics, Seattle, Washington, 45–54. https://doi.org/10.18653/v1/2022.socialnlp-1.4

[Sample21] Leo Hemamou and William Coleman. 2022. Delivering Fairness in Human Resources AI: Mutual Information to the Rescue. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (Eds.). Association for Computational Linguistics, Online only, 867–882. https://aclanthology.org/2022.aacl-main.64

[Sample22] Samhita Honnavalli, Aesha Parekh, Lily Ou, Sophie Groenwold, Sharon Levy, Vicente Ordonez, and William Yang Wang. 2022. Towards Understanding Gender-Seniority Compound Bias in Natural Language Generation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 1665–1670. https://aclanthology.org/2022.lrec-1.177

[Sample23] Silke Husse and Andreas Spitz. 2022. Mind Your Bias: A Critical Review of Bias Detection Methods for Contextual Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 4212–4234. https://doi.org/10.18653/v1/2022.findings-emnlp.311

[Sample24] Sophie Jentzsch and Cigdem Turan. 2022. Gender Bias in BERT - Measuring and Analysing Biases through Sentiment Rating in a Realistic Downstream Classification Task. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, Christian Hardmeier, Christine Basta,

Marta R. Costa-jussà, Gabriel Stanovsky, and Hila Gonen (Eds.). Association for Computational Linguistics, Seattle, Washington, 184–199. https://doi.org/10.18653/v1/2022.gebnlp-1.20

[Sample25] Alexander Kwako, Yixin Wan, Jieyu Zhao, Kai-Wei Chang, Li Cai, and Mark Hansen. 2022. Using Item Response Theory to Measure Gender and Racial Bias of a BERT-based Automated English Speech Assessment System. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch (Eds.). Association for Computational Linguistics, Seattle, Washington, 1–7. https://doi.org/10.18653/v1/2022.bea-1.1

[Sample26] Bum Chul Kwon and Nandana Mihindukulasooriya. 2022. An Empirical Study on Pseudo-log-likelihood Bias Measures for Masked Language Models Using Paraphrased Sentences. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, Apurv Verma, Yada Pruksachatkun, Kai-Wei Chang, Aram Galstyan, Jwala Dhamala, and Yang Trista Cao (Eds.). Association for Computational Linguistics, Seattle, U.S.A., 74–79. https://doi.org/10.18653/v1/2022.trustnlp-1.7

[Sample27] Nayeon Lee, Yejin Bang, Tiezheng Yu, Andrea Madotto, and Pascale Fung. 2022a. NeuS: Neutral Multi-News Summarization for Mitigating Framing Bias. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 3131–3148. https://doi.org/10.18653/v1/2022.naacl-main.228

[Sample28] Young-Jun Lee, Chae-Gyun Lim, Yunsu Choi, Ji-Hui Lm, and Ho-Jin Choi. 2022b. PERSONACHATGEN: Generating Personalized Dialogues using GPT-3. In *Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge*, Heuiseok Lim, Seungryong Kim, Yeonsoo Lee, Steve Lin, Paul Hongsuck Seo, Yumin Suh, Yoonna Jang, Jungwoo Lim, Yuna Hur, and Suhyune Son (Eds.). Association for Computational Linguistics, Gyeongju, Republic of Korea, 29–48. https://aclanthology.org/2022.ccgpk-1.4

[Sample29] Yizhi Li, Ge Zhang, Bohao Yang, Chenghua Lin, Anton Ragni, Shi Wang, and Jie Fu. 2022. HERB: Measuring Hierarchical Regional Bias in Pre-trained Language Models. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (Eds.). Association for Computational Linguistics, Online only, 334–346. https://aclanthology.org/2022.findings-aacl.32

[Sample30] Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. Socially Aware Bias Measurements for Hindi Language Representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 1041–1052. https://doi.org/10.18653/v1/2022.naacl-main.76

[Sample31] Ninareh Mehrabi, Umang Gupta, Fred Morstatter, Greg Ver Steeg, and Aram Galstyan. 2022. Attributing Fair Decisions with Attention Interventions. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, Apurv Verma, Yada Pruksachatkun, Kai-Wei Chang, Aram Galstyan, Jwala Dhamala, and Yang Trista Cao (Eds.). Association for Computational Linguistics, Seattle, U.S.A., 12–25. https://doi.org/10.18653/v1/2022.trustnlp-1.2

[Sample32] Isar Nejadgholi, Esma Balkir, Kathleen Fraser, and Svetlana Kiritchenko. 2022. Towards Procedural Fairness: Uncovering Biases in How a Toxic Language Classifier Uses Sentiment Information. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, Jasmijn Bastings, Yonatan Belinkov, Yanai Elazar, Dieuwke Hupkes, Naomi Saphra, and Sarah Wiegreffe (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 225–237. https://doi.org/10.18653/v1/2022.blackboxnlp-1.18

[Sample33] Yoona Park and Frank Rudzicz. 2022. Detoxifying Language Models with a Toxic Corpus. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, Bharathi Raja Chakravarthi, B Bharathi, John P McCrae, Manel Zarrouk, Kalika Bali, and Paul Buitelaar (Eds.). Association for Computational Linguistics, Dublin, Ireland, 41–46. https://doi.org/10.18653/v1/2022.ltedi-1.6

[Sample34] Victor Petren Bach Hansen, Atula Tejaswi Neerkaje, Ramit Sawhney, Lucie Flek, and Anders Sogaard. 2022. The Impact of Differential Privacy on Group Disparity Mitigation. In *Proceedings of the Fourth Workshop on Privacy in Natural Language Processing*, Oluwaseyi Feyisetan, Sepideh Ghanavati, Patricia Thaine, Ivan Habernal, and Fatemehsadat Mireshghallah (Eds.). Association for Computational Linguistics, Seattle, United States, 12–12. https://doi.org/10.18653/v1/2022.privatenlp-1.2

[Sample35] Rajkumar Pujari, Erik Oveson, Priyanka Kulkarni, and Elnaz Nouri. 2022. Reinforcement Guided Multi-Task Learning Framework for Low-Resource

Stereotype Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 6703–6712. https://doi.org/10.18653/v1/2022.acl-long.462

[Sample36] Alan Ramponi and Sara Tonelli. 2022. Features or Spurious Artifacts? Data-centric Baselines for Fair and Robust Hate Speech Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 3027–3040. https://doi.org/10.18653/v1/2022.naacl-main.221

[Sample37] Danielle Saunders, Rosie Sallis, and Bill Byrne. 2022. First the Worst: Finding Better Gender Translations During Beam Search. In *Findings of the Association for Computational Linguistics: ACL 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 3814–3823. https://doi.org/10.18653/v1/2022.findings-acl.301

[Sample38] Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2022. Does Representational Fairness Imply Empirical Fairness?. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (Eds.). Association for Computational Linguistics, Online only, 81–95. https://aclanthology.org/2022.findings-aacl.8

[Sample39] Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 9180–9211. https://doi.org/10.18653/v1/2022.emnlp-main.625

[Sample40] Anders Søgaard. 2022. Should We Ban English NLP for a Year?. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 5254–5260. https://doi.org/10.18653/v1/2022.emnlp-main.351

[Sample41] Victor Steinborn, Philipp Dufter, Haris Jabbar, and Hinrich Schuetze. 2022. An Information-Theoretic Approach and Dataset for Probing Gender Stereotypes in Multilingual Masked Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 921–932. https://doi.org/10.18653/v1/2022.findings-naacl.69

[Sample42] Tianxiang Sun, Junliang He, Xipeng Qiu, and Xuanjing Huang. 2022. BERTScore is Unfair: On Social Bias in Language Model-Based Metrics for Text Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3726–3739. https://doi.org/10.18653/v1/2022.emnlp-main.245

[Sample43] Terne Sasha Thorn Jakobsen, Maria Barrett, Anders Søgaard, and David Lassen. 2022. The Sensitivity of Annotator Bias to Task Definitions in Argument Mining. In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, Sameer Pradhan and Sandra Kuebler (Eds.). European Language Resources Association, Marseille, France, 44–61. https://aclanthology.org/2022.law-1.6

[Sample44] Eddie Ungless, Amy Rafferty, Hrichika Nag, and Björn Ross. 2022. A Robust Bias Mitigation Procedure Based on the Stereotype Content Model. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, David Bamman, Dirk Hovy, David Jurgens, Katherine Keith, Brendan O'Connor, and Svitlana Volkova (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 207–217. https://doi.org/10.18653/v1/2022.nlpcss-1.23

[Sample45] Jialu Wang, Yang Liu, and Xin Wang. 2022. Assessing Multilingual Fairness in Pre-trained Multimodal Representations. In *Findings of the Association for Computational Linguistics: ACL 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 2681–2695. https://doi.org/10.18653/v1/2022.findings-acl.211

[Sample46] Genta Winata, Shijie Wu, Mayank Kulkarni, Thamar Solorio, and Daniel Preotiuc-Pietro. 2022. Cross-lingual Few-Shot Learning on Unseen Languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Yulan He, Heng Ji,

Sujian Li, Yang Liu, and Chua-Hui Chang (Eds.). Association for Computational Linguistics, Online only, 777–791. https://aclanthology.org/2022.aacl-main.59

[Sample47] Guillaume Wisniewski, Lichao Zhu, Nicolas Ballier, and François Yvon. 2022. Analyzing Gender Translation Errors to Identify Information Flows between the Encoder and Decoder of a NMT System. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, Jasmijn Bastings, Yonatan Belinkov, Yanai Elazar, Dieuwke Hupkes, Naomi Saphra, and Sarah Wiegreffe (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 153–163. https://doi.org/10.18653/v1/2022.blackboxnlp-1.13

[Sample48] Jinrui Yang, Sheilla Njoto, Marc Cheong, Leah Ruppanner, and Lea Frermann. 2022. Professional Presentation and Projected Power: A Case Study of Implicit Gender Information in English CVs. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, David Bamman, Dirk Hovy, David Jurgens, Katherine Keith, Brendan O'Connor, and Svitlana Volkova (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 140–150. https://doi.org/10.18653/v1/2022.nlpcss-1.15

[Sample49] Shiyue Zhang, Vishrav Chaudhary, Naman Goyal, James Cross, Guillaume Wenzek, Mohit Bansal, and Francisco Guzman. 2022a. How Robust is Neural Machine Translation to Language Imbalance in Multilingual Tokenizer Training?. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, Kevin Duh and Francisco Guzmán (Eds.). Association for Machine Translation in the Americas, Orlando, USA, 97–116. https://aclanthology.org/2022.amta-research.8

[Sample50] Zhexin Zhang, Jiale Cheng, Hao Sun, Jiawen Deng, Fei Mi, Yasheng Wang, Lifeng Shang, and Minlie Huang. 2022b. Constructing Highly Inductive Contexts for Dialogue Safety through Controllable Reverse Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3684–3697. https://doi.org/10.18653/v1/2022.findings-emnlp.270

Table of labels for the first research question: Ethical grounding

Label group	Labels
Paper-level labels	
wrongness not discussed	
EXAMPLES: This is a paper-level label, so there are no concrete quotes that are labeled. But to illustrate, a paper in this group would only define the subject (bias) like in this quote from [Sample43] :"We define bias as an inclination or prejudice for or against something, e.g. groups, individuals, concepts and behaviors. The term social bias can be used in two senses: an individual's bias which is explained by the (social) group the individual belongs to, and bias against (social) groups. The latter is typically the focus of bias studies in NLP ... .". There is no clear statement that bias is negative; it is just described as an inclination in attitude or opinion.	
Quote-level labels	
no ethical vocabulary	negative impact, undesirable behavior, undesired bias/stereotypes, unsafe behavior, violate the law
EXAMPLES: "Societal biases in NLP has raised increasing attention because large-scale LMs containing societal biases can produce undesirable biased expressions and have negative societal impacts on the minorities ..." [Sample29] , "In recent years, there has been a series of works aiming to measure social biases or other unwanted behaviors in NLP." [Sample39]	
ethical grounding in terms of harm	allocational harm, environmental harm, generate harmful language, harmful bias/stereotypes, representational harm, unspecific harm/ harmful behavior
EXAMPLES: "Our results hint at how zero-shot generalization may provide some hopeful representation toward minimizing harm and bias in these large-scale language models." [Sample4] , "The word representations that are derived from these models (so-called word embeddings) also retain potentially harmful biases contained in the data that are used in the training process ... ." [Sample23]	
other ethical vocabulary	lack of accountability, lack of transparency, no equitable NLG/NLP, toxic behavior, unfair behavior, untrustworthy AI
EXAMPLES: " language models (LMs) that empower much of the web as we know it are well known to contain biases that promote structural discrimination in downstream tasks against minorities and larger social groups alike ... ." [Sample23] , "recent studies criticize that the multilingual textual representations do not learn equally high-quality representations for all the languages ... ." [Sample45]	
Second order quote-level labels	
development of concepts	developed, not developed
EXAMPLES: " If there exists a correlation between a certain gender and an attribute, the language model intrinsically and perpetually causes representational harm Blodgett et al. (2020) through improper preconceptions." [Sample2] is labeled as developed; "... LMs propagate stereotypical behavior which could be toxic towards any demographic group (more towards disadvantaged groups)." [Sample1] is labeled as not developed, because it is not made explicit how propagating stereotypical behavior can be toxic.	
grounded in theoretical literature	theoretical grounding, no theoretical grounding
EXAMPLES: The first quote from the last examples provides theoretical grounding, the second does not. The criteria here is whether literature is cited that uses theoretical work from social science, political sciences, philosophy or neighboring disciplines.	

Figure 5: This table gives an overview of the labels used to annotate for the first research question. We distinguish three types of labels: (i) paper-level labels can only given to the paper as a whole, (ii) quote-level labels are given to quotes in papers, consequently, papers can have multiple labels of this type in the same category, (iii) second-order quote-level labels are labels that need to be attached to quote-level labels within that category.

Table of labels for the second research question:
Causes of wrongful behavior of AI systems

Label group	Labels
Causes of wrongful behavior of AI systems	
issues with training	biased/limited training data
EXAMPLES: "Potential harms arise when biases around word choice or grammatical gender inflections reflect demographic or social biases (Sun et al., 2019)." [Sample37] .	
technological causes	bias in AI system, tech fix, impaired performance
EXAMPLES: " If there exists a correlation between a certain gender and an attribute, the language model intrinsically and perpetually causes representational harm (Blodgett et al., 2020) through improper preconceptions." [Sample2] is an example for *bias in AI system*. "Differential privacy may often hurt minority groups the most, and reducing the fairness gap by focusing on minority groups during training typically puts their privacy at risk." [Sample34] for *tech fix*.	
issues with development	lack of investment in minorities, methodologically weak research, unsupervised automation
EXAMPLES: "Languages without large corpora also often face a lack of support in common NLP tools and unexpected errors in other tools due to a lack of testing and use." [Sample7] exemplifies *lack of investment in minorities*. "... researchers' unsystematic approach to experimental design when measuring bias in language generation is a symptom of 1) under-utilizing domain-expert, human annotators to annotate bias generation validation sets and 2) undermining the importance of certain settings that are usually deemed incidental for other NLP tasks, which in fact could be pivotal in bias measurement task." [Sample3] was labeled as *methodologically weak research*.	
issues with use	naive use of biased systems
EXAMPLES: "... naive use of ML methods bears the risk of exacerbating bias" [Sample48] .	
Causes of bias in AI system	
issues with training	annotator bias, biased/limited training data
EXAMPLES: "... human-generated data used to build these systems is considered the main source of these biases" [Sample5] .	
technological causes	limited model capacity, loss function, model properties
EXAMPLES: "We reveal the capacity of the model and cross-entropy loss in knowledge distillation have a negative effect on social bias." [Sample2] , "These examples convey that biases can be hidden in the attention mechanism, and thus pose the risk of being recovered in representations and predictions." [Sample19] .	
issues with development	lack of investment in minorities, subjective decisions, (unsupervised) automation
EXAMPLES: "We find that biases evaluation is extremely sensitive to different design choices when curating template prompts." [Sample6] for *subjective decision*.	
Causes of issues with training	
societal bias	
EXAMPLES: "... these systems learn a wide variety of societal biases embedded in human text ..." [Sample8] .	

Figure 6: This table gives an overview of the labels used to annotate for the second research question. For this question we did not use paper-level labels or second-order labels. However, because of the central role bias played in many papers we labeled not only for the cause of wrong, but, if bias was mentioned as such cause, we also labeled the cause of bias. We used for both levels of causes the same grouping of labels: issues with training, technological causes, issues with development. Issues with use could not occur as cause of bias in an AI system. Finally, to give a complete picture of the causal network that was considered, we also labeled causes given for the causes of system bias. There was only one relevant group of causes considered: societal biases.

Table of labels for the third research question: Proposed solutions

Label group	Labels
Quote-level labels	
technological solutions	technological solutions, + transparent AI, + bias measurement
EXAMPLES: "In this work, we specifically investigate how robust NMT is to language imbalance in tokenizer training." [Sample49] . "Explainability techniques (XAI) can potentially help in discovering and quantifying biases." [Sample32] for *+ transparent AI*, "It is important to perform a systematic, robust and automated bias analysis to help build equitable NLG systems." [Sample1] for *+ bias measure*.	
solutions targeting the researchers/developers	diversify approach, create awareness, improve conceptualization, provide dataset statements, improve social/cultural awareness of researchers/developers, study context of use, fix problem before launch, improve annotation, report environmental costs
EXAMPLES: "Extending the KG to other languages as well as data sources could yield a more global view on stereotypes regarding a specific entity." [Sample17] for *diversify approach*, "To raise awareness and mitigate stereotypical reflection, we need to understand how biases emerge and how they are reinforced." [Sample24] for *create awareness*, "We propose that an artifacts statement should be added to this documentation as a way to contribute to diagnosis (and thus mitigation) of pre-existing bias, which is also one of the goals of data statements." for *dataset statements*.	
solutions targeting institutionsor infrastructure	policies/guidelines, research/development agenda
EXAMPLES: "Concretely, we should strive to establish robust estimators of bias, clean and curated test sets, and guidelines for their rigorous applications within the (often restrictive) confines of language model APIs to avoid measuring the biases that we introduce in the process of detecting them" [Sample23] , "This makes it imperative that associated harms be understood not just for the Western world and with a focus on English language models, but also across languages and cultures." [Sample30] .	
involving other agents in development	interdisciplinary approach, crowdsourcing, domain experts for fairness, community efforts
EXAMPLES: "At the same time, we argue that addressing biases in language models requires and deserves a concerted community effort ..." [Sample23] , "The usage of our fair classifiers in the real world should be carefully monitored with the aid of domain experts." [Sample38] .	
others	provide guidance for users, advise industry practitioners
EXAMPLES: "Furthermore, there is a lack of research advising on how industry practitioners can effectively and cheaply debias outputs whilst retaining quality, accuracy and realism.", [Sample13] , "Understanding the risk of uncontrolled adoption of such automatic tools, careful guidance should be provided in how to adopt them." [Sample27] .	
Second order quote-level labels	
target solution	target solution

Figure 7: This table gives an overview of the labels used to annotate for the third research question. For this question we used quote-level labels and one second order label: we labeled all solutions discussed in the paper and then marked the solution the paper actually worked on in the paper with an additional label. In the group *technological solutions* we added additional labels for solutions that target transparent AI and bias measurement.