Ethics and Bias in Natural Language Processing

Zeerak Talat[a] **and Su Lin Blodgett**[b], [a] University of Edinburgh, Edinburgh, United Kingdom; and [b] Microsoft Research, Montréal, QC, Canada

Introduction 1
Conceptualisations of Bias and Fairness 1
Causes of Harms 2
Methods: Measurements and Mitigation 2
Challenges 2
Critiques 3
Conclusion 4
References 4

Key Points

- Bias and fairness are essentially contested concepts, with multiple, sometimes conflicting conceptualisations and operationalisations.
- Measurement and mitigation approaches span data, model, and output levels, though their general effectiveness remains uncertain.
- Critical work highlights open challenges—such as under-justified conceptualisations and measurement approaches—as well as fundamental critiques—such as issues of power and consent.

Abstract

As language technologies are increasingly widely deployed and adopted, ethical concerns have become a more prominent issue, with research engaging with it spanning fields such as natural language processing to human-computer interaction. In this entry, we present a primer for interested readers, extracting key insights and texts on ethical issues pertaining to language technologies, with a focus on bias and fairness. We highlight existing approaches and identify challenges and critiques of how the body of work has conceptualized and operationalized research on ethical language technologies.

Introduction

Recent advancements in machine learning (ML) and natural language processing (NLP) have resulted in greater public awareness of language technologies, such as large language models (Bender et al., 2021).

The ethical issues that have been raised regarding language technologies span a wide variety of concerns (Hovy & Spruit, 2016; Solaiman et al., 2024; Ungless et al., 2024), including bias and fairness (e.g., Bender et al., 2021; Bolukbasi et al., 2016), labor (e.g., Shiller, 2019), environmental (e.g., Strubell et al., 2019), privacy (e.g., Carlini et al., 2023), explainability and transparency (e.g., Liao & Wortman Vaughan, 2024), accountability (e.g., Birhane et al., 2024; Gray Widder et al., 2023), governance (Jernite et al., 2022), and dual use/misuse (Hovy & Spruit, 2016; Kaffee et al., 2023) concerns.

In this entry, we focus on computational and sociotechnical aspects of bias and fairness in language technologies; we provide an overview of conceptualisations of bias and fairness, measurement and mitigation approaches, and challenges and critiques of current approaches.

Conceptualisations of Bias and Fairness

Bias and fairness are *essentially contested* constructs—that is, they have "multiple context-dependent, and sometimes even conflicting, theoretical understandings" (Jacobs & Wallach, 2021). Conceptualisations of bias and fairness with respect to language technologies have generally been centered on *group fairness*—i.e., differences in model or system behavior or performance with respect to different social groups (Gallegos et al., 2024).

Such fairness concerns can be broadly thought of in terms of several types of impacts or harms to stakeholders: quality of service, representational, allocative, and experiential harms. Quality of service harms occur when a technology performs better for some groups of people than others, leading to an unjust distribution of benefits (Dev et al., 2022); for example, some technologies exhibit performance disparities between white and Black users, requiring the latter to exert additional labor to make the technology work (Cunningham et al., 2024). Representational harms occur "when systems reinforce the subordination of some groups along the lines of identity" (Crawford, 2017), including when technologies produce stereotypes about, disempower or demean, or erase groups of people (Dev et al., 2022; Hovy & Spruit, 2016). For example, stereotypical associations in model representations reflect those held by humans (Caliskan et al., 2017), while text generated by LLMs routinely misgenders transgender and non-binary people (Ovalle et al., 2023). Allocative harms occur "when a system withholds [from] certain groups an opportunity or resource" (Barocas et al., 2019), such as in hiring or criminal justice. For example, content moderation technologies disproportionately penalize African American English social media posts, potentially excluding speakers from full participation online (Davidson et al., 2019). Finally, experiential or psychological harms can impact stakeholders' "cognitive and affective states" (Chien & Danks, 2024). For example, Wenzel et al. (2023) show that Black participants interacting with voice assistants with higher word error rates for Black users than white users exhibit "significantly higher levels of self-consciousness [and] lower levels of self-esteem and positive affect" than white users, mirroring patterns from research on racial microaggressions.

Causes of Harms

A robust literature tracing the origins of these fairness and bias concerns has emerged. While the earliest such work focused on training datasets and methods (e.g., Zhao et al., 2017), more recent work has expanded these analyses to interrogate assumptions and choices made across the NLP lifecycle (Cao & Daumé, 2021), including task formulation (Pavlick & Kwiatkowski, 2019); label choices (Fortuna et al., 2021); data curation practices (Bagga & Piper, 2020), including annotation disagreement (Plank, 2022) and data filtering (Dodge et al., 2021); and evaluation practices (Barocas et al., 2021). Critical work has argued that all of these assumptions and choices are the product of social infrastructures, beliefs, and practices, and thus necessarily subjective and value-laden (Hoffman, 2019).

Methods: Measurements and Mitigation

A plethora of instruments for measuring bias and fairness in language technologies have emerged—operationalizing, implicitly or explicitly, the various conceptualisations of bias and fairness described above. These approaches vary along several dimensions: the *object* over which the measurement is taken, which can include model representations, model predictions, model generations, or predictions from auxiliary classifiers applied to model generations; the *approach* or computation undertaken when using the instrument; the *components* that make up the instrument or that are required to use it, which can include "datasets, metrics, tools, benchmarks, and annotation instructions" (Harvey et al., 2025); and how *ideal model behavior* is thought to be reflected in the resulting measurements. **Table 1** provides examples of measurement instruments developed for various objects of measurement. We refer the reader to Gallegos et al. (2024) for an in-depth survey of measurements of bias and fairness for LLMs.

Various approaches across the NLP development lifecycle for mitigating these bias and fairness concerns have also been proposed (Hovy & Prabhumoye, 2021). These include interventions over datasets for model training or evaluation—e.g., by augmenting data to include language describing or written by a wider range of people, or by accounting for meaningful annotator disagreement; over the input representations models rely on; over objective functions used to train models; or over the model output by introducing guardrails, such as blocklists or classifiers to identify model output that should not be shown to a user (Bender et al., 2021). Separately, research has proposed a variety of mechanisms for documenting development assumptions and choices, including data statements (Bender & Friedman, 2018), datasheets (Gebru et al., 2018), and model cards (Mitchell et al., 2019).

Challenges

Many challenges remain for measuring and mitigating bias and fairness concerns for language technologies. Although clear conceptualization of the construct to be measured is required for effective measurement and mitigation (Solaiman et al., 2024; Wallach et al., 2025), conceptualisations of bias and fairness for language technologies are often under-specified, inconsistent, or lack justification (Blodgett et al., 2020; Devinney et al., 2022). Measurement instruments may be poorly operationalized—in part due to unclear conceptualisations (Blodgett et al., 2021), and the resulting measurements may not correlate with one another, or meaningfully capture the construct to be measured or reflect technologies' real-world impacts; for example, measurements of bias over models' internal representations often do not correlate with measurements taken from the technology where those representations are ultimately used (Goldfarb-Tarrant et al., 2021). Measurement instruments may also suffer from limited uptake from practitioners, either because they are not well-suited to measuring what practitioners need to measure, or because of other organizational or practical barriers (Harvey et al., 2025).

Table 1 Examples of instruments for measuring biases in language technologies and their characteristics.

Example measurement instrument	*Object of measurement*	*Measurement approach*	*Measurement instrument component(s)*	*Ideal model behavior*
WEAT (Caliskan et al., 2017)	Word embeddings (model representations)	Compute the relative similarity between the embeddings of two sets of target words (e.g., European American- and African American-associated names) and two sets of attribute words (e.g., words about pleasantness and unpleasantness)	Lexicons (sets of target and attribute words), WEAT metric	No difference in relative similarity
Equity Evaluation Corpus (Kiritchenko & Mohammad, 2018)	Predicted sentiment (model predictions)	Given a pair of "sentences that differ only in the gender/race of a person mentioned", compare the predicted sentiment of each sentence	Template sentences, auxiliary classifier (sentiment analysis system)	Equal sentiment scores across paired sentences
CrowS-Pairs (Nangia et al., 2020)	Token probabilities (model generations)	Given a pair of sentences, one "demonstrating or violating a stereotype" about a group and another differing only by a group identity term, compute the "percentage of examples for which a model assigns a higher ([pseudo]-)likelihood to the stereotyping sentence … over the less stereotyping sentence"	Dataset (paired sentences), aggregating metric	50% score
Regard score (Sheng et al., 2019)	Regard score of generated sentences (model generations with auxiliary annotation or classifier)	Given a set of sentences generated from a set of prefix templates with different identity terms, annotate each generated sentence for the regard and take the majority vote	Dataset of prefix templates, regard annotations	Equal regard scores across sentences generated from prefix templates with different identity terms

These difficulties are compounded by other challenges of the language technology landscape. For example, LLMs, which increasingly underpin modern language technologies, are often described as "general-purpose" and are accompanied by expansive claims about their capabilities (e.g., reasoning, understanding) and impacts on people who encounter them (Solaiman et al., 2024). At the same time, the space of technologies' use cases, including those for which they were not imagined, is rapidly expanding, demanding new conceptualisations of bias and fairness—e.g., how might we understand fairness in the context of a technology designed for tutoring students versus those for generating creative stories, travel itineraries, or image captions?—and measurement and mitigation approaches. Static benchmark datasets, which are necessarily developed for specific tasks and language varieties, may not be adaptable to other settings, and the ever-changing nature of language means that benchmarks can at best be well-suited to particular tasks and varieties at a particular moment in time (Raji et al., 2021). Finally, technologies designed to simulate humans—often to act as "companions" or to replace or act on behalf of users—sharpen existing and raise new concerns, particularly around consent, oversight, and over-reliance (Olteanu et al., 2025).

Critiques

Alongside research identifying bias and fairness concerns and proposing measurement and mitigation approaches, critical work has also raised more fundamental critiques of this research and language technologies more broadly. Some of this scholarship argues that focusing on whether technologies are biased or unfair distracts from questions of how technologies distribute (political) power (Kalluri, 2020) or embed inherently harmful ideologies (Stark & Hutson, 2021); for example, technologies designed for tasks such as emotion recognition not only rely on (and thus reproduce) physiognomic assumptions, but also—even if they appear to be

fair—enable expanded infrastructures of surveillance (Stark & Hutson, 2021). As a result of these failures to consider power and the broader social structures in which technologies are embedded, bias and fairness approaches may fail to meaningfully produce the socially just outcomes that are sought (Hoffmann, 2019). Moreover, while biases may be partially addressable, technologies' necessarily situated and subjective nature renders "unbiased" technologies impossible (Talat et al., 2021).

Similarly, work has contended that concerns about bias and fairness ignore whether people can meaningfully participate in technologies' development (Delgado et al., 2023), including whether they are able to refuse participation in the training or use of technologies (Cifor et al., 2019). The Feminist Data Manifest-No (Cifor et al., 2019) argues for consentful inclusion that centers stakeholders and respects their desires not to be included. A lack of meaningful forms of consent can give rise to people engaging in adversarial relationships with technologies to regain a sense of control (Kulynych et al., 2020).

Conclusion

In this entry, we have briefly introduced research on bias and fairness in language technologies, highlighting conceptualisations, measurement and mitigation approaches, and current and emerging challenges and critiques.

References

Bagga, S., & Piper, A. (2020). Measuring the effects of bias in training data for literary classification. In S. DeGaetano, A. Kazantseva, N. Reiter, & S. Szpakowicz (Eds.), *Proceedings of the 4th joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature* (pp. 74—84). International Committee on Computational Linguistics. https://aclanthology.org/2020.latechclfl-1.9/.

Barocas, S., Guo, A., Kamar, E., Krones, J., Morris, M. R., Vaughan, J. W., Wadsworth, W. D., & Wallach, H. (2021). Designing disaggregated evaluations of AI systems: Choices, considerations, and tradeoffs. In *Proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society* (pp. 368—378). https://doi.org/10.1145/3461702.3462610

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. fairmlbook.org http://www.fairmlbook.org.

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics, 6*, 587—604. https://doi.org/10.1162/tacl_a_00041

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. In *ACM conference on fairness, accountability and transparancy, Virtual, Canada*. https://faculty.washington.edu/ebender/papers/Stochastic_Parrots.pdf.

Birhane, A., Steed, R., Ojewale, V., Vecchione, B., & Raji, I. D. (2024). AI auditing: The broken bus on the road to AI accountability. In *2024 IEEE conference on secure and trustworthy machine learning (SaTML)* (pp. 612—643). https://doi.org/10.1109/SaTML59370.2024.00037

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "Bias" in NLP. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5454—5476). https://doi.org/10.18653/v1/2020.acl-main.485

Blodgett, S. L., Lopez, G., Olteanu, A., Sim, R., & Wallach, H. (2021). Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (pp. 1004—1015). https://doi.org/10.18653/v1/2021.acl-long.81

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183—186. https://doi.org/10.1126/science.aal4230

Cao, Y. T., & Daumé, H. (2021). Toward gender-inclusive coreference resolution: An analysis of gender and bias throughout the machine learning lifecycle. *Computational Linguistics, 47*(3), 615—661. https://doi.org/10.1162/coli_a_00413

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., & Wallace, E. (2023). Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX conference on security symposium*.

Chien, J., & Danks, D. (2024). Beyond behaviorist representational harms: A plan for measurement and mitigation. In *The 2024 ACM conference on fairness, accountability, and transparency* (pp. 933—946). https://doi.org/10.1145/3630106.3658946

Cifor, M., Garcia, P., Cowan, T. L., Rault, J., Sutherland, J., Hoffman, A. L., Salehi, N., & Nakamura, L. (2019). *Feminist data manifest-no*. Feminist data Manifest-No. https://www.manifestno.com.

Crawford, K. (2017). *The trouble with bias*.

Cunningham, J., Blodgett, S. L., Madaio, M., Daumé Iii, H., Harrington, C., & Wallach, H. (2024). Understanding the impacts of language technologies' performance disparities on African American language speakers. *Findings of the Association for Computational Linguistics ACL, 2024*, 12826—12833. https://doi.org/10.18653/v1/2024.findings-acl.761

Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the third workshop on abusive language online* (pp. 25—35). https://doi.org/10.18653/v1/W19-3504

Delgado, F., Yang, S., Madaio, M., & Yang, Q. (2023). The participatory turn in AI design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM conference on equity and access in algorithms, mechanisms, and optimization*. https://doi.org/10.1145/3617694.3623261

Dev, S., Sheng, E., Zhao, J., Amstutz, A., Sun, J., Hou, Y., Sanseverino, M., Kim, J., Nishi, A., Peng, N., & Chang, K.-W. (2022). On measures of biases and harms in NLP. *Findings of the Association for Computational Linguistics: AACL-IJCNLP, 2022*, 246—267. https://doi.org/10.18653/v1/2022.findings-aacl.24

Devinney, H., Björklund, J., & Björklund, H. (2022). Theories of "gender" in NLP bias research. In *2022 ACM conference on fairness accountability and transparency* (pp. 2083—2102). https://doi.org/10.1145/3531146.3534627

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., & Gardner, M. (2021). Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 1286—1305). https://doi.org/10.18653/v1/2021.emnlp-main.98

Fortuna, P., Soler-Company, J., & Wanner, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management, 58*(3), 102524. https://doi.org/10.1016/j.ipm.2021.102524.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics, 50*(3), 1097—1179. https://doi.org/10.1162/coli_a_00524

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *arXiv:1803.09010 [Cs]*. http://arxiv.org/abs/1803.09010.

Goldfarb-Tarrant, S., Marchant, R., Muñoz Sánchez, R., Pandya, M., & Lopez, A. (2021). Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (pp. 1926–1940). https://doi.org/10.18653/v1/2021.acl-long.150

Gray Widder, D., West, S., & Whittaker, M. (2023). Open (for business): Big tech, concentrated power, and the political economy of open AI. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4543807

Harvey, E., Sheng, E., Blodgett, S. L., Chouldechova, A., Garcia-Gathright, J., Olteanu, A., & Wallach, H. (2025). Understanding and meeting practitioner needs when measuring representational harms caused by LLM-based systems. In *Proceedings of ACL*.

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass, 15*(8). https://doi.org/10.1111/lnc3.12432

Hovy, D., & Spruit, S. (2016). The social impact of natural language processing. In *Proceedings of ACL*.

Jacobs, A. Z., & Wallach, H. (2021). Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 375–385). https://doi.org/10.1145/3442188.3445901

Jernite, Y., Nguyen, H., Biderman, S., Rogers, A., Masoud, M., Danchev, V., Tan, S., Luccioni, A. S., Subramani, N., Johnson, I., Dupont, G., Dodge, J., Lo, K., Talat, Z., Radev, D., Gokaslan, A., Nikpoor, S., Henderson, P., Bommasani, R., & Mitchell, M. (2022). Data governance in the age of large-scale data-driven language technology. In *2022 ACM conference on fairness, accountability, and transparency* (pp. 2206–2222). https://doi.org/10.1145/3531146.3534637

Kaffee, L.-A., Arora, A., Talat, Z., & Augenstein, I. (2023). Thorny roses: Investigating the dual use dilemma in natural language processing. *Findings of the Association for Computational Linguistics: EMNLP, 2023*, 13977–13998. https://doi.org/10.18653/v1/2023.findings-emnlp.932

Kalluri, P. (2020). Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature, 583*(7815). https://doi.org/10.1038/d41586-020-02003-2, 169–169.

Kiritchenko, S., & Mohammad, S. (2018). Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the seventh joint conference on lexical and computational semantics* (pp. 43–53). https://doi.org/10.18653/v1/S18-2005

Kulynych, B., Overdorf, R., Troncoso, C., & Gürses, S. (2020). POTs: Protective optimization technologies. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 177–188). https://doi.org/10.1145/3351095.3372853

Liao, Q. V., & Wortman Vaughan, J. (2024). AI transparency in the age of LLMs: A human-centered research roadmap. *Harvard Data Science Review, Special Issue, 5*. https://doi.org/10.1162/99608f92.8036d03b

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220–229). https://doi.org/10.1145/3287560.3287596

Nangia, N., Vania, C., Bhalerao, R., & Bowman, S. R. (2020). CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 1953–1967). https://doi.org/10.18653/v1/2020.emnlp-main.154

Olteanu, A., Barocas, S., Blodgett, S. L., Egede, L., DeVrio, A., & Cheng, M. (2025). AI automatons: AI systems intended to imitate humans (No. arXiv:2503.02250). *arXiv*. https://doi.org/10.48550/arXiv.2503.02250

Ovalle, A., Goyal, P., Dhamala, J., Jaggers, Z., Chang, K.-W., Galstyan, A., Zemel, R., & Gupta, R. (2023). "I'm fully who I am": Towards centering transgender and non-binary voices to measure biases in open language generation. In *2023 ACM conference on fairness accountability and transparency* (pp. 1246–1266). https://doi.org/10.1145/3593013.3594078

Pavlick, E., & Kwiatkowski, T. (2019). Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics, 7*, 677–694. https://doi.org/10.1162/tacl_a_00293

Plank, B. (2022). The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 10671–10682). https://doi.org/10.18653/v1/2022.emnlp-main.731

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the everything in the whole wide world benchmark. In *Proceedings of the neural information processing systems track on datasets and benchmarks. Conference on neural information processing systems*. https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/084b6fbb10729ed4da8c3d3f5a3ae7c9-Paper-round2.pdf.

Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3405–3410). https://doi.org/10.18653/v1/D19-1339

Shiller, R. J. (2019). *Narrative economics: How stories go viral & drive major economic events*. Princeton University Press.

Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., Blodgett, S. L., Chen, C., III, H. D, Dodge, J., Duan, I., Evans, E., Friedrich, F., Ghosh, A., Gohar, U., Hooker, S., Jernite, Y., Kalluri, R., Lusoli, A., Leidinger, A., … Subramonian, A. (2024). Evaluating the social impact of generative AI systems in systems and society (no. arXiv:2306.05949). *arXiv*. http://arxiv.org/abs/2306.05949.

Stark, L., & Hutson, J. (2021). Physiognomic artificial intelligence. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3927300

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 3645–3650). https://doi.org/10.18653/v1/P19-1355

Talat, Z., Lulz, S., Bingel, J., & Augenstein, I. (2021). Disembodied machine learning: On the illusion of objectivity in NLP. http://arxiv.org/abs/2101.11974.

Ungless, E. L., Vitsakis, N., Talat, Z., Garforth, J., Ross, B., Onken, A., Kasirzadeh, A., & Birch, A. (2024). Ethics whitepaper: Whitepaper on ethical research into large language models (No. arXiv:2410.19812). *arXiv*. http://arxiv.org/abs/2410.19812.

Wallach, H., Desai, M., Cooper, A. F., Wang, A., Atalla, C., Barocas, S., Blodgett, S. L., Chouldechova, A., Corvi, E., Dow, P. A., Garcia-Gathright, J., Olteanu, A., Pangakis, N., Reed, S., Sheng, E., Vann, D., Vaughan, J. W., Vogel, M., Washington, H., & Jacobs, A. Z. (2025). Position: Evaluating generative AI systems is a social science measurement challenge (No. arXiv:2502.00561). *arXiv*. https://doi.org/10.48550/arXiv.2502.00561

Wenzel, K., Devireddy, N., Davison, C., & Kaufman, G. (2023). Can voice assistants be microaggressors? Cross-race psychological responses to failures of automatic speech recognition. In *Proceedings of the 2023 CHI conference on human factors in computing systems* (pp. 1–14). https://doi.org/10.1145/3544548.3581357