Gaps Between Research and Practice When Measuring Representational Harms Caused by LLM-Based Systems

Emma Harvey[1*] **Emily Sheng**[2] **Su Lin Blodgett**[2] **Alexandra Chouldechova**[2]
Jean Garcia-Gathright[2] **Alexandra Olteanu**[2] **Hanna Wallach**[2]
[1]Cornell University [2]Microsoft Research
evh29@cornell.edu

1 Introduction

As large language model (LLM)-based systems become increasingly widespread, so too has their potential to cause *representational harms* [3, 18], which occur when a system "represents some social groups in a less favorable light than others, demeans them, or fails to recognize their existence altogether" [8]. Representational harms are abstract concepts that cannot be measured directly [38]—yet measuring such harms is important, as they can cause tangible negative outcomes, e.g., through the entrenchment of harmful social hierarchies, which may affect people's belief systems and psychological states [17, 62, 16]. To facilitate the measurement of representational harms, the NLP research community has produced and made publicly available numerous *measurement instruments*,[1] including tools [e.g., 40, 14], datasets [e.g., 64, 28, 32, 34, 56], metrics [e.g., 11, 15, 10, 58, 43], benchmarks (consisting of both datasets and metrics) [e.g., 23, 48, 45, 46, 55, 63, 22, 24, 26, 27, 31], annotation instructions [e.g., 42], and other techniques [e.g., 36, 52, 61]. However, the research community lacks clarity about whether and to what extent these instruments meet the needs of practitioners tasked with developing and deploying LLM-based systems in the real world, and how the instruments could be improved.

Via a series of semi-structured interviews with practitioners (N = 12) in a variety of roles in different organizations, we identify four types of challenges that prevent practitioners from effectively using publicly available instruments for measuring representational harms caused by LLM-based systems: (1) challenges related to *using publicly available measurement instruments*; (2) challenges related to doing measurement *in practice*; (3) challenges arising from measurement tasks *involving LLM-based systems*; and (4) challenges specific to measuring *representational harms*. Our goal is to advance the development of instruments for measuring representational harms that are well-suited to practitioner needs, thus better facilitating the responsible development and deployment of LLM-based systems.

2 Methods

We conducted 12 semi-structured interviews with practitioners who reported that their work involved measuring representational harms caused by LLM-based systems. We recruited participants through our professional networks, social media, cold emails to individuals identified through LinkedIn and conference proceedings, and snowball sampling [44]. All interviews were one-hour long and were conducted virtually between June and August 2024. Participants provided informed consent before their interviews and received $75 gift cards afterwards. The study was approved by our institution's IRB.

Participants. Participants held research (7/12), applied science (2/12), engineering (2/12), and consulting (1/12) roles at large tech companies (6/12), AI-focused startups (3/12), large non-tech companies (2/12), and AI-focused nonprofits (1/12). They described working on a variety of LLM-based systems, including search engines and chatbots, as well as on content moderation tools for LLMs.

Interviews and analysis. We asked participants to describe their roles and the LLM-based systems they worked on. We also asked them to walk us through an example of how they measured representational

[*]Work conducted during an internship at Microsoft Research

[1]By *publicly available*, we mean that an instrument has been made available on the internet or via an academic publication for others to use or adapt, potentially subject to licensing considerations.

NeurIPS 2024 Workshop on Evaluating Evaluations (EvalEval).

harms, noting the publicly available measurement instruments they used or considered using. We then asked them to reflect on their experiences with those instruments and to discuss any challenges. Although our sample size is relatively small (a common problem when conducting research on AI practitioners [57]), we conducted interviews until we reached saturation, i.e., until multiple consecutive interviews did not uncover any new perspectives [59, 35]. Finally, we analyzed the resulting interview transcripts using a thematic analysis with an inductive–deductive coding approach [12, 13].

3 Results and Discussion

We identified four types of challenges that prevent practitioners from effectively using publicly available instruments for measuring representational harms caused by LLM-based systems.

Challenges related to using publicly available measurement instruments. Although prior work has identified a range of challenges related to using publicly available measurement instruments [e.g., 9, 60], participants primarily reported challenges related to *validity* and *specificity*. Almost all participants (11/12) mentioned issues of validity—i.e., whether a measurement instrument meaningfully measures what stakeholders think it measures [38]—related to *correctness* (e.g., tools produce inaccurate outputs, datasets contain mislabeled instances) and *contestedness* (e.g., different instruments use different definitions of representational harms). Similarly, almost all participants (11/12) mentioned issues of specificity—i.e., whether a measurement instrument is sufficiently specific to a system, its use case(s), and its deployment context(s). Examples included datasets that are too generic to align with relevant use cases like customer service chats, and labels that are not sufficiently detailed (e.g., categories like "hate speech" are labeled, but details like the targeted social group are missing).

Challenges related to doing measurement in practice. Particularly salient to our participants (and in line with considerable prior work [e.g., 2, 6, 7, 19–21, 29, 37, 39, 41, 47, 49–51, 53]) were challenges related to doing measurement *in practice*—i.e., when working within the constraints surrounding products and services that are deployed to real users. Some participants (4/12) felt that their measurements needed to produce *specific quality assurances*, and suggested that those assurances might be better achieved through software testing practices. They also pointed to data licensing and security issues (3/12), the need to align with company-specific policies (2/12), a lack of time to find publicly available measurement instruments in the first place (2/12), and competitive pressures[2] (1/12) as factors incentivizing them to build new measurement instruments rather than adopting existing ones.

Challenges arising from measurement tasks involving LLM-based systems. Because the datasets that are used to train LLMs are often unknown, it can be difficult to tell whether an LLM-based system that performs well on a benchmark has simply been trained using the benchmark data. Multiple participants (6/12) therefore expressed discomfort with using publicly available benchmarks and datasets, even if they are valid and specific to their needs. Some reported finding suitable publicly available benchmarks, but then using them only as inspiration to create new, internal benchmarks from scratch.

Challenges specific to measuring representational harms. Finally, participants distinguished representational harms from other types of harms. Specifically, many (9/12) felt that compared to other types of harms (e.g., privacy violations), representational harms required more and different information to measure. Participants felt that measuring representational harms required context (6/12), alignment on essentially constructed constructs (2/12), and social science expertise (2/12) that measuring other types of harms did not. Some participants (2/12) also felt that they faced less commercial incentive to measure representational harms compared to other types of harms (e.g., quality of service harms), causing them to limit their measurement efforts. For example, some participants reported that relevant stakeholders felt that aligned models were so unlikely to generate outright demeaning content that it was not worth measuring in the first place. Others found that if they did not know how to *mitigate* certain representational harms, relevant stakeholders would not value the measurement.

The four types of challenges described above shed light on whether and to what extent publicly available instruments for measuring representational harms meet the needs of practitioners tasked with developing and deploying LLM-based systems in the real world. Future work should further investigate these types of challenges and take steps to address them by, for example, drawing on measurement theory from the social sciences [e.g., 1, 38] and pragmatic measurement [30, 33] to *improve instruments* for measuring representational harms, and on other fields like implementation science [e.g., 5, 4, 25, 54] to *improve the uptake* of publicly available measurement instruments among practitioners.

[2]Specifically, pressures to develop new measurement instruments in order to claim proprietary capabilities.

Broader Impacts

By making explicit the key challenges that prevent practitioners from effectively using publicly available instruments for measuring representational harms caused by LLM-based systems, our work is intended to serve as a starting point to *bridge gaps between research and practice*. We hope that our findings will serve as a foundation for future work on measurement instruments that are better suited to practitioners' needs, as well as work on the adoption of publicly available measurement instruments.

Limitations

The primary limitation of our study is our small sample size. As is the case with many studies targeting technology workers [57], it was challenging to identify and recruit potential participants. Practitioners who held relevant roles often declined to speak with us due to NDAs or other confidentiality concerns. As a result, we were only able to interview 12 practitioners, some of whom declined to answer certain questions in order to remain in compliance with their employers' NDAs.

The page limit did not allow us to meaningfully discuss ways to bridge gaps between research and practice by addressing the challenges we identified. We plan to expand on this in future work by drawing on measurement theory, pragmatic measurement, and implementation science to identify potential ways to improve instruments for measuring representational harms and their uptake among practitioners.

References

[1] Robert Adcock and David Collier. Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *American Political Science Review*, 95 (3):529–546, September 2001. ISSN 0003-0554, 1537-5943. doi: 10.1017/S0003055401003100. URL https://www.cambridge.org/core/product/identifier/S0003055401003100/type/journal_article.

[2] Agathe Balayn, Mireia Yurrita, Jie Yang, and Ujwal Gadiraju. "Fairness Toolkits, A Checkbox Culture?" On the Factors that Fragment Developer Practices in Handling Algorithmic Harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES, pages 482–495, Montreal QC Canada, August 2023. ACM. ISBN 9798400702310. doi: 10.1145/3600211.3604674. URL https://dl.acm.org/doi/10.1145/3600211.3604674.

[3] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. The problem with bias: Allocative versus representational harms in machine learning. In *Proceedings of SIGCIS*, Philadelphia, PA, 2017.

[4] Mark S. Bauer and JoAnn Kirchner. Implementation science: What is it and why should I care? *Psychiatry Research*, 283:112376, January 2020. ISSN 01651781. doi: 10.1016/j.psychres.2019.04.025. URL https://linkinghub.elsevier.com/retrieve/pii/S016517811930602X.

[5] Mark S. Bauer, Laura Damschroder, Hildi Hagedorn, Jeffrey Smith, and Amy M. Kilbourne. An introduction to implementation science for the non-specialist. *BMC Psychology*, 3(1): 32, December 2015. ISSN 2050-7283. doi: 10.1186/s40359-015-0089-9. URL http://bmcpsychology.biomedcentral.com/articles/10.1186/s40359-015-0089-9.

[6] Glen Berman, Nitesh Goyal, and Michael Madaio. A Scoping Study of Evaluation Practices for Responsible AI Tools: Steps Towards Effectiveness Evaluations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI, pages 1–24, Honolulu HI USA, May 2024. ACM. ISBN 9798400703300. doi: 10.1145/3613904.3642398. URL https://dl.acm.org/doi/10.1145/3613904.3642398.

[7] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H. Chi. Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES, pages 453–459, New York, NY, USA, January 2019. Association for Computing Machinery. ISBN 978-1-4503-6324-2. doi: 10.1145/3306618.3314234. URL https://dl.acm.org/doi/10.1145/3306618.3314234.

[8] Su Lin Blodgett, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://www.aclweb.org/anthology/2020.acl-main.485.

[9] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 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)*, pages 1004–1015, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.81. URL https://aclanthology.org/2021.acl-long.81.

[10] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, NIPS, December 2016. URL https://papers.nips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html. arXiv:1607.06520 [cs, stat].

[11] Rishi Bommasani and Percy Liang. Trustworthy Social Bias Measurement, 2022. URL https://arxiv.org/abs/2212.11672.

[12] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, January 2006. ISSN 1478-0887, 1478-0895. doi: 10.1191/1478088706qp063oa. URL http://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa.

[13] Virginia Braun and Victoria Clarke. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, 11(4):589–597, August 2019. ISSN 2159-676X, 2159-6778. doi: 10.1080/2159676X.2019.1628806. URL https://www.tandfonline.com/doi/full/10.1080/2159676X.2019.1628806.

[14] Zana Buçinca, Chau Minh Pham, Maurice Jakesch, Marco Tulio Ribeiro, Alexandra Olteanu, and Saleema Amershi. AHA!: Facilitating AI Impact Assessment by Generating Examples of Harms, 2023. URL https://arxiv.org/abs/2306.03280.

[15] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, April 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aal4230. URL https://www.science.org/doi/10.1126/science.aal4230.

[16] Jennifer Chien and David Danks. Beyond Behaviorist Representational Harms: A Plan for Measurement and Mitigation. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 933–946, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi: 10.1145/3630106.3658946. URL https://dl.acm.org/doi/10.1145/3630106.3658946.

[17] Emily Corvi, Hannah Washington, Stefanie Reed, Chad Atalla, Alexandra Chouldechova, Alex Dow, Jean Garcia-Gathright, Nicholas Pangakis, Emily Sheng, Dan Vann, Matthew Vogel, and Hanna Wallach. Representational harms through the lens of speech act theory. Unpublished manuscript, 2024.

[18] Kate Crawford. The trouble with bias. In *Conference on Neural Information Processing Systems (invited speaker)*, 2017.

[19] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT, pages 473–484, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533113. URL https://dl.acm.org/doi/10.1145/3531146.3533113.

[20] Wesley Hanwen Deng, Boyuan Guo, Alicia Devrio, Hong Shen, Motahhare Eslami, and Kenneth Holstein. Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18, Hamburg Germany, April 2023. ACM. ISBN 978-1-4503-9421-5. doi: 10.1145/3544548.3581026. URL https://dl.acm.org/doi/10.1145/3544548.3581026.

[21] Wesley Hanwen Deng, Nur Yildirim, Monica Chang, Motahhare Eslami, Kenneth Holstein, and Michael Madaio. Investigating Practices and Opportunities for Cross-functional Collaboration around AI Fairness in Industry Practice. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 705–716, Chicago IL USA, June 2023. ACM. ISBN 9798400701924. doi: 10.1145/3593013.3594037. URL https://dl.acm.org/doi/10.1145/3593013.3594037.

[22] Sunipa Dev, Akshita Jha, Jaya Goyal, Dinesh Tewari, Shachi Dave, and Vinodkumar Prabhakaran. Building Stereotype Repositories with Complementary Approaches for Scale and Depth. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 84–90, Dubrovnik, Croatia, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.c3nlp-1.9. URL https://aclanthology.org/2023.c3nlp-1.9.

[23] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT, pages 862–872, Virtual Event Canada, March 2021. ACM. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445924. URL https://dl.acm.org/doi/10.1145/3442188.3445924.

[24] Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. Multi-Dimensional Gender Bias Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.23. URL https://www.aclweb.org/anthology/2020.emnlp-main.23.

[25] M.W. Enkin and A.R. Jadad. Using anecdotal information in evidence-based health care: Heresy or necessity? *Annals of Oncology*, 9(9):963–966, September 1998. ISSN 09237534. doi: 10.1023/A:1008495101125. URL https://linkinghub.elsevier.com/retrieve/pii/S0923753419481714.

[26] David Esiobu, Xiaoqing Tan, Saghar Hosseini, Megan Ung, Yuchen Zhang, Jude Fernandes, Jane Dwivedi-Yu, Eleonora Presani, Adina Williams, and Eric Smith. ROBBIE: Robust Bias Evaluation of Large Generative Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 3764–3814, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.230. URL https://aclanthology.org/2023.emnlp-main.230.

[27] Eve Fleisig, Aubrie Amstutz, Chad Atalla, Su Lin Blodgett, Hal Daumé Iii, Alexandra Olteanu, Emily Sheng, Dan Vann, and Hanna Wallach. FairPrism: Evaluating Fairness-Related Harms in Text Generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6231–6251, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.343. URL https://aclanthology.org/2023.acl-long.343.

[28] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.301. URL https://www.aclweb.org/anthology/2020.findings-emnlp.301.

[29] Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. *Journal of Artificial Intelligence Research*, 77:103–166, May 2023. ISSN 1076-9757. doi: 10.1613/jair.1.13715. URL https://www.jair.org/index.php/jair/article/view/13715.

[30] Russell E. Glasgow and William T. Riley. Pragmatic Measures. *American Journal of Preventive Medicine*, 45(2):237–243, August 2013. ISSN 07493797. doi: 10.1016/j.amepre.2013.03.010. URL https://linkinghub.elsevier.com/retrieve/pii/S0749379713002651.

[31] Hila Gonen and Kellie Webster. Automatically Identifying Gender Issues in Machine Translation using Perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1991–1995, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.180. URL https://www.aclweb.org/anthology/2020.findings-emnlp.180.

[32] Rishav Hada, Agrima Seth, Harshita Diddee, and Kalika Bali. "Fifty Shades of Bias": Normative Ratings of Gender Bias in GPT Generated English Text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 1862–1876, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.115. URL https://aclanthology.org/2023.emnlp-main.115.

[33] David J Hand. *Measurement: A very short introduction*. Oxford University Press, 2016.

[34] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.234. URL https://aclanthology.org/2022.acl-long.234.

[35] Monique Hennink and Bonnie N. Kaiser. Sample sizes for saturation in qualitative research: A systematic review of empirical tests. *Social Science & Medicine*, 292:114523, January 2022. ISSN 02779536. doi: 10.1016/j.socscimed.2021.114523.

[36] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. Ai generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028):147–154, Sep 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07856-5. URL https://doi.org/10.1038/s41586-024-07856-5.

[37] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI, pages 1–16, Glasgow Scotland Uk, May 2019. ACM. ISBN 978-1-4503-5970-2. doi: 10.1145/3290605.3300830. URL https://dl.acm.org/doi/10.1145/3290605.3300830.

[38] Abigail Z. Jacobs and Hanna Wallach. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT, pages 375–385, New York, NY, USA, March 2021. Association for Computing Machinery. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445901. URL https://dl.acm.org/doi/10.1145/3442188.3445901.

[39] Michelle Seng Ah Lee and Jat Singh. The Landscape and Gaps in Open Source Fairness Toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI, pages 1–13, Yokohama Japan, May 2021. ACM. ISBN 978-1-4503-8096-6. doi: 10.1145/3411764.3445261. URL https://dl.acm.org/doi/10.1145/3411764.3445261.

[40] Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3197–3207, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3539147. URL https://doi.org/10.1145/3534678.3539147.

[41] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI, pages 1–14, Honolulu HI USA, April 2020. ACM. ISBN 978-1-4503-6708-0. doi: 10.1145/3313831.3376445. URL https://dl.acm.org/doi/10.1145/3313831.3376445.

[42] Ahmed Magooda, Alec Helyar, Kyle Jackson, David Sullivan, Chad Atalla, Emily Sheng, Dan Vann, Richard Edgar, Hamid Palangi, Roman Lutz, Hongliang Kong, Vincent Yun, Eslam Kamal, Federico Zarfati, Hanna Wallach, Sarah Bird, and Mei Chen. A Framework for Automated Measurement of Responsible AI Harms in Generative AI Applications, October 2023. URL http://arxiv.org/abs/2310.17750. arXiv:2310.17750 [cs].

[43] Gaurav Maheshwari, Aurélien Bellet, Pascal Denis, and Mikaela Keller. Fair Without Leveling Down: A New Intersectional Fairness Definition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 9018–9032, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.558. URL https://aclanthology.org/2023.emnlp-main.558.

[44] David L Morgan. Snowball sampling. *The SAGE encyclopedia of qualitative research methods*, 2:815–16, 2008.

[45] Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.416. URL https://aclanthology.org/2021.acl-long.416.

[46] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.154. URL https://www.aclweb.org/anthology/2020.emnlp-main.154.

[47] Victor Ojewale, Ryan Steed, Briana Vecchione, Abeba Birhane, and Inioluwa Deborah Raji. Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling, February 2024. URL http://arxiv.org/abs/2402.17861.

[48] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.165. URL https://aclanthology.org/2022.findings-acl.165.

[49] Joaquin Quiñonero Candela, Yuwen Wu, Brian Hsu, Sakshi Jain, Jennifer Ramos, Jon Adams, Robert Hallman, and Kinjal Basu. Disentangling and Operationalizing AI Fairness at LinkedIn. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT, pages 1213–1228, New York, NY, USA, June 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594075. URL https://dl.acm.org/doi/10.1145/3593013.3594075.

[50] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT*, pages 33–44, New York, NY, USA, January 2020. Association for Computing Machinery. ISBN 978-1-4503-6936-7. doi: 10.1145/3351095.3372873. URL https://dl.acm.org/doi/10.1145/3351095.3372873.

[51] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–23, April 2021. ISSN 2573-0142. doi: 10.1145/3449081. URL https://dl.acm.org/doi/10.1145/3449081.

[52] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July 2020. Association for Computational

Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442.

[53] Brianna Richardson, Jean Garcia-Gathright, Samuel F. Way, Jennifer Thom, and Henriette Cramer. Towards Fairness in Practice: A Practitioner-Oriented Rubric for Evaluating Fair ML Toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Yokohama Japan, May 2021. ACM. ISBN 978-1-4503-8096-6. doi: 10.1145/3411764.3445604. URL https://dl.acm.org/doi/10.1145/3411764.3445604.

[54] Everett M. Rogers. *Diffusion of innovations*. Free Press [u.a.], New York, NY, 3. ed edition, 1962. ISBN 978-0-02-926650-2.

[55] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2002. URL http://aclweb.org/anthology/N18-2002.

[56] Sandra Sandoval, Jieyu Zhao, Marine Carpuat, and Hal Daumé Iii. A Rose by Any Other Name would not Smell as Sweet: Social Bias in Names Mistranslation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 3933–3945, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.239. URL https://aclanthology.org/2023.emnlp-main.239.

[57] Morgan Klaus Scheuerman. In the Walled Garden: Challenges and Opportunities for Research on the Practices of the AI Tech Industry. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 456–466, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi: 10.1145/3630106.3658918. URL https://dl.acm.org/doi/10.1145/3630106.3658918.

[58] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3405–3410, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1339. URL https://www.aclweb.org/anthology/D19-1339.

[59] Mario Luis Small. 'how many cases do i need?': On science and the logic of case selection in field-based research. *Ethnography*, 10(1):5–38, 2009. doi: 10.1177/1466138108099586.

[60] Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Canyu Chen, Hal Daumé III au2, Jesse Dodge, Isabella Duan, Ellie Evans, Felix Friedrich, Avijit Ghosh, Usman Gohar, Sara Hooker, Yacine Jernite, Ria Kalluri, Alberto Lusoli, Alina Leidinger, Michelle Lin, Xiuzhu Lin, Sasha Luccioni, Jennifer Mickel, Margaret Mitchell, Jessica Newman, Anaelia Ovalle, Marie-Therese Png, Shubham Singh, Andrew Strait, Lukas Struppek, and Arjun Subramonian. Evaluating the social impact of generative ai systems in systems and society, 2024. URL https://arxiv.org/abs/2306.05949.

[61] Angelina Wang, Solon Barocas, Kristen Laird, and Hanna Wallach. Measuring Representational Harms in Image Captioning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT, pages 324–335, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533099. URL https://dl.acm.org/doi/10.1145/3531146.3533099.

[62] Angelina Wang, Xuechunzi Bai, Solon Barocas, and Su Lin Blodgett. Measuring stereotype harm from machine learning errors requires understanding who is being harmed by which errors in what ways. In *ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*, 2023.

[63] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, Volume 2 (Short Papers), pages 15–20, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL http://aclweb.org/anthology/N18-2003.

[64] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bl8u7ZRlbM.

A Appendix

Our semi-structured interview guide is shown below. As is typical of a semi-structured interview process, not every participant was asked exactly the same questions in exactly the same order, and some participants were asked additional follow-up or clarifying questions based on the answers they provided. The interview questions were supplemented with a set of slides containing definitions of key terms that we screenshared with participants. The definitions are included in the script below.

A.1 Introductions [5 min]

Welcome! Thank you so much for taking the time for this interview. Before we get started, I just want to quickly introduce myself, talk about the goals of this study, and give you a chance to ask any questions you might have. This research study is intended to understand gaps between research and practice in evaluating large language model (LLM)-based systems, with a focus on measuring harms, adverse impacts, or other undesirable behaviors. In this interview, I'll ask you to share your experiences with and opinions on such evaluations, without discussing confidential information. I will also record this interview for the purpose of creating a deidentified transcript. If you prefer that your video not be recorded, please feel free to turn your camera off at this time. In addition, if at any point you would like to skip a question, take a break, or end the interview, please feel free to do so.

Do you have any questions before we get started?

A.2 Background [5 min]

[IQ1] To start, please briefly describe your role, focusing on your professional experience as it relates to LLM-based systems.

[IQ2] Can you briefly describe the LLM-based system(s) that you have previously evaluated, currently evaluate, or plan to evaluate?

A.3 Experience with measurement instruments for representational harms [15 min]

[IQ3] Throughout this interview, I will be focusing primarily on representational harms, which occur when "a system represents some social groups in a less favorable light than it represents other groups by stereotyping them, demeaning them, or failing to recognize their existence altogether."

What examples of representational harms caused by LLM-based systems are you aware of?

If interviewee was not familiar with representational harms, we provided the following examples:

- LLMs might reinforce stereotypes, for example, by using the word "nurse" to refer to a female healthcare provider and the word "doctor" to refer to a male healthcare provider in otherwise identical contexts.
- LLMs might generate slurs or derogatory language about a social group.
- LLMs might erase a social group, for example, by only listing male athletes when a user asks for examples of talented soccer players, thus failing to recognize the existence of non-male soccer players.

[IQ4] Do your previous, current, or planned evaluation(s) of LLM-based system(s) involve measuring representational harms?

[IQ5] What types of representational harms are you measuring?

[IQ6] Can you walk me through, from start to finish, an example of how you measure representational harms? I'm especially interested in hearing about how you decided on your approach, whether

you relied on existing, publicly available tools, benchmarks, datasets, metrics, annotation guidelines, and so on, or whether you decided to develop your own.

To allow for open-ended discussion, we did not provide participants with a specific definition of 'measurement instruments'; rather, we provided the following examples of instruments:

- An example of a tool is Perspective API.
- An example of a benchmark is StereoSet, which includes a dataset of prompts that could elicit stereotyping content with corresponding metrics that measure the extent to which a language model produces stereotypes.
- An example of a dataset is WildChat, which is a corpus of 1 million real user-ChatGPT interactions.
- Examples of metrics are the Word and Sentence Embedding Association Tests (WEAT and SEAT), which measure whether "attribute words" (e.g. male, female) are disproportionately associated with a set of "target words" (e.g. different professions).
- Annotation instructions are sets of instructions and examples for humans to use when annotating system outputs for particular properties.
- An example of another type of instrument is Matched Guide Probing, a method adapted from sociolinguistics.

For each instrument mentioned, we asked the following questions:

[IQ7] What type(s) of representational harms are you measuring with [this instrument]?

[IQ8] How did you decide to use [this instrument]?

[IQ9] How do you use [this instrument] in your evaluation(s)?

[IQ10] Where did [this instrument] come from? Did you develop it yourself, modify an existing [instrument], or use an existing [instrument] as-is?

If applicable, for one instrument that the interviewee developed themselves, we asked the following questions:

[IQ11] Why did you decide to develop [this instrument] yourself?

[IQ12] What, if any, actions have you taken or plan to take upon seeing the measurements obtained using [this instrument]?

If applicable, for one instrument that the interviewee adapted from an existing instrument, we asked the following questions:

[IQ13] Why did you decide to start with this existing [instrument]?

[IQ14] Why did you decide to modify [this instrument] rather than using it as-is?

[IQ15] What, if any, actions have you taken or plan to take upon seeing the measurements obtained using [this instrument]?

If applicable, for one instrument that the interviewee used as-is, we asked the following questions:

[IQ16] Why did you decide to use this existing [instrument] as-is?

[IQ17] What, if any, actions have you taken or plan to take upon seeing the measurements obtained using [this instrument]?

A.4 Challenges with measurement instruments for representational harms [15 min]

[IQ18] Were there any other existing, publicly available [instruments] that you investigated using instead?

[IQ19] *For each instrument mentioned:* Why did you decide not to use [this instrument]?

[IQ20] *For each of the challenges defined below, say either:*

"It sounds like you mentioned an issue to do with [challenge]. Is that correct?", *or*

"I don't think you mentioned [challenge]. Did you experience any issues with this?"

We provided interviewees with the following set of challenges related to measurement instruments:

- Whether it is sufficiently specific to the system being evaluated and its particular use cases and deployment contexts
- Whether it can be adapted for different systems, use cases, and deployment contexts
- Whether it results in valid measurements – i.e., meaningfully measures what stakeholders think it measures
- Whether it results in similar measurements when used in similar ways, especially over time
- Whether its resulting measurements can be understood by stakeholders
- Whether its resulting measurements can be acted upon by stakeholders
- Whether it can scale to increasing workloads

[IQ21] *For each challenge experienced:* What, if anything, did you do to address this issue?

[IQ22] Did you experience any other issues that we haven't discussed?

[IQ23] *If applicable:* What, if anything, did you do to address this issue?

A.5 Comparing measurement of representational harms to other harms [5 min]

[IQ24] Do your previous, current, or planned evaluation(s) of LLM-based system(s) involve measuring harms, adverse impacts, or other undesirable behaviors other than representational harms?

[IQ25] *If yes:* What types of harms, adverse impacts, or other undesirable behaviors?

[IQ26] *If yes:* Are your experiences measuring these types of harms, adverse impacts, or other undesirable behaviors similar to your experiences measuring representational harms? What, if anything, is similar and what, if anything, is different about your experiences? I'm especially interested in hearing about how the [instruments] you use to measure these types of harms, adverse impacts, or other undesirable behaviors are similar to or different from the [instruments] you use to measure representational harms.

A.6 Desired improvements to measuring representational harms [5 min]

[IQ27] Putting aside any time or budget constraints, what, if anything, would you improve about the way that you previously, currently, or plan to measure representational harms?

[IQ28] What do you need, that you don't currently have, in order to make those improvements?

A.7 Closing [5 min]

[IQ29] Is there anything else you would like to tell us about your previous, current, or planned evaluation(s) of LLM-based system(s)?