Patterns

CellPress
OPEN ACCESS

Perspective

Benchmark suites instead of leaderboards for evaluating AI fairness

Angelina Wang,[1,2,*] Aaron Hertzmann,[3] and Olga Russakovsky[1]
[1]Princeton University, Princeton, NJ, USA
[2]Stanford University, Stanford, CA, USA
[3]Adobe Research, San Francisco, CA, USA
*Correspondence: angelina.wang@stanford.edu
https://doi.org/10.1016/j.patter.2024.101080

THE BIGGER PICTURE As artificial intelligence (AI) systems become increasingly complex and ubiquitous, developers and users need to understand these systems' social fairness impacts: their potential for harm to different communities. Fairness is a debated concept, and constructing measurements of social fairness is challenging. Doing so requires considering different applications, capturing different types of harms, and reflecting different communities' perspectives. Hence, the common approach of scoring models by AI fairness leaderboards, which attempt to reduce fairness to a single number, is inappropriate. Instead of measuring single scores, we argue for comprehensive benchmark suites, which expose multiple measurements, allowing researchers and practitioners to better understand the different trade-offs between AI models, without hiding potential harms within a single score. We describe principles for the construction and use of such suites.

SUMMARY

Benchmarks and leaderboards are commonly used to track the fairness impacts of artificial intelligence (AI) models. Many critics argue against this practice, since it incentivizes optimizing for metrics in an attempt to build the "most fair" AI model. However, this is an inherently impossible task since different applications have different considerations. While we agree with the critiques against leaderboards, we believe that the use of benchmarks can be reformed. Thus far, the critiques of leaderboards and benchmarks have become unhelpfully entangled. However, benchmarks, when not used for leaderboards, offer important tools for understanding a model. We advocate for collecting benchmarks into carefully curated "benchmark suites," which can provide researchers and practitioners with tools for understanding the wide range of potential harms and trade-offs among different aspects of fairness. We describe the research needed to build these benchmark suites so that they can better assess different usage modalities, cover potential harms, and reflect diverse perspectives. By moving away from leaderboards and instead thoughtfully designing and compiling benchmark suites, we can better monitor and improve the fairness impacts of AI technology.

INTRODUCTION

How do we understand the potential social harms and inequities of artificial intelligence (AI) models?[1–3] One way that AI fairness researchers have responded to this question is by creating benchmarks. Each benchmark includes a dataset and associated metric,[4] intending to measure a particular dimension of AI fairness. As an example, the Bias Benchmark for QA (BBQ) uses multiple-choice questions to measure language models' bias toward different demographic groups by querying the model with questions such as about whether a boy or a girl more likely struggles with math.[5]

A common practice when developing new algorithms is to report scores on leaderboards. A leaderboard ranks algorithms based on a single score that is either computed over one benchmark or aggregated over multiple benchmarks. These rankings exist on dynamic websites determining the state-of-the-art,[6–9] as well as in the tables of academic papers to establish superiority over prior methods. Technical research agendas sometimes focus on building AI models that top a leaderboard.

Many recent authors criticize the use of benchmarks and leaderboards within fairness research.[10–12] They rightly point out that different harms cannot be meaningfully averaged into a single score, and moreover, benchmark optimization encourages

"gaming" the metrics. These points echo concerns across AI in general regarding the use of benchmarks and leaderboards.[4,13–18]

Nonetheless, benchmarks can be quite useful for identifying potential harms and operationalizing ethical principles.[19] Indeed, some of the most impactful fairness changes in AI models have come from quantitative measurements (e.g., Gender Shades,[20] COMPAS[21]). Arguably, many of the criticisms of benchmarks are really criticisms of leaderboards. Can we use benchmarks productively without leaderboards?

We argue that fairness benchmarks should be used not for leaderboards, but rather collected into "benchmark suites" that provide insights into a model's trade-offs around a set of potential harms. Benchmark suites, as we call them, are collections of individual benchmarks, consisting of multiple datasets and metrics. While these collections are sometimes referred to as benchmarks as well, we reserve the term "benchmark" for individual measurements. Recent examples for large language models (LLMs) that fall under our definition of a benchmark suite include DecodingTrust,[22] TrustLLM,[23] and HELM,[24] as well as the set of evaluations used in model cards like Claude 3's[25] and Llama's.[26] However, despite being collections of benchmarks, these collections do not meet what we will present as best practices for benchmark suites.

We propose that a suite should be curated with the goal of completeness in capturing the construct of interest. Suites can then serve as a guiding tool for selecting among all of the proliferating benchmarks. The results of applying these suites can provide researchers and practitioners ways to analyze the behavior of their models, anticipate potential problem areas, and understand the corresponding trade-offs. For example, consider the task of understanding the ability of an LLM to select merit-based scholarship recipients. One LLM might make a race-blind decision and another may take a race-aware approach. Separate benchmark metrics can expose this trade-off, while other metrics might reveal effects relating to gender and intersectional dynamics. The preferred model, if any at all, will depend on the deployment scenario, for example, whether the scholarship is diversity based or whether race-blind legislation applies to the scholarship. A benchmark suite here would help to proactively identify potential discrimination harms and trade-offs between models, without determining one model to be categorically better than another.

We need to understand, however, how this idea fits within the standard paradigm of technical AI research, which focuses on optimizing for and comparing models against leaderboards. Here, we present constructive ways that benchmark suites can be used in lieu of leaderboards toward the goals of making research progress and selecting models for deployment.

Making benchmark suites effective for these uses requires further research to improve them. We suggest three necessary improvements toward this goal: creating benchmarks specific to individual modalities, systematically analyzing harm coverage and limitations, and representing diverse perspectives. In doing so, we put forth benchmark suite curation as a critical research direction.

Overall in this work, we focus on examples and recommendations specifically attuned for the fairness setting, but many of these arguments generalize to measurement in AI at large.

THE PROBLEM WITH FAIRNESS LEADERBOARDS

Many have criticized the use of leaderboards to rank fairness algorithms.[13,14,27] Leaderboards for a single benchmark misleadingly use one measurement to capture the entire notion of fairness. Even leaderboards for a collection of benchmarks report rankings based on a single score aggregated over the whole collection, so that an algorithm may give an improved "fairness score" by increasing performance on one benchmark while making it worse on another. However, one benchmark might be assessing a very different potential harm from another; for some applications, this apparent improvement could actually be much worse.

For example, an image captioning algorithm can be used for different applications, such as indexing image search results, or assisting blind and low-vision users. Some blind and low-vision users have expressed a desire for detailed image descriptions that may include inferred gender because that provides useful social context for understanding an image.[28] However, transgender and non-binary individuals may object to automated gender classification because of the dignitary harms of being misgendered or the reification of rigid gender categories.[29,30] A leaderboard averaging two different measurements, one around accurate gender inference and one around the lack of any gender inference, would obscure these tensions and be counterproductive.

This points to a fundamental issue: whether one model is preferred over another depends on the context in which it is used and the ethical values that are prioritized in that context.[31] Hence, during research, different fairness considerations cannot be meaningfully summarized with a single score, making leaderboards misleading. Pareto improvements (i.e., the notion that one model can, in all respects, match or outperform another model) will often be impossible under well-constructed suites of fairness benchmarks because some measurements will directly conflict. A related scenario can be seen in the famous fairness impossibility theorems, which demonstrate that in most real-world settings of binary classification such as criminal risk prediction, it is impossible to satisfy two reasonable measures of fairness (e.g., equal error rates and calibrated predictions) simultaneously.[32,33]

Even when considering one metric at a time, pursuing metrics can create serious problems. Optimization relies on each metric to be a perfect measure of the unobservable fairness construct being measured.[34] This metric optimization can then cause problems such as overemphasizing easily measured short-term goals (e.g., clickthrough rate over reported preference)[16] or leveling-down models (i.e., achieving fairness by making most groups worse off).[35] Optimization is also likely to push models into edge cases where measurement becomes misdirected (e.g., Goodhart's law).[36] For example, a benchmark that productively measures vulgar language, when optimized, may remove innocuous uses of language where individuals reclaim certain slurs.[37]

For these reasons, fairness leaderboards should be abandoned. This, however, does not mean abandoning fairness benchmarks, especially benchmark suites, which can provide a comprehensive and meaningful assessment of AI fairness without reducing findings to a single leaderboard number.

BENCHMARK SUITES FOR ASSESSING FAIRNESS

We argue for the development and use of benchmark suites. Benchmark suites are evolving collections of benchmarks specific to a modality (e.g., computer vision, natural language processing) and particular task that can vary in specificity (e.g., chatbot, algebra tutor chatbot). Application context may change both the composition of the benchmark suite and/or the interpretation of a standardized suite. Suites may be curated or may emerge through common practice and consensus. The validity of benchmark suites should then be assessed not only at the benchmark level but suite level as well. We recommend the use of "suite" instead of "benchmark" to describe these collections, since this conflation in terminology has made it easy to expect one score from a suite and neglect multi-faceted coverage.

Benchmark suites should be accompanied by tools to help visualize and interpret the measurement results. It will likely be important to look into the actual inputs and outputs of a model during this interpretation.[24] Researchers and practitioners alike can then use these results to assess the prevalence and strength of a set of fairness harms for new models and algorithms. These findings can be used to guide research progress as well as model selection for deployment.

Making research progress

Leaderboards have been fundamental to making research progress. Without comparing methods based on leaderboard performance, we need a new way to judge progress. We propose progress to take the form of new methods that control trade-offs that we did not have the levers to access before. This would expand the available fairness tools that practitioners can choose from, and in determining which kinds of levers are more valuable, also motivate research progress based on work that grounds fairness in concrete applications.[38,39] In our proposal, a fairness researcher fine-tuning an LLM can run a suite of fairness benchmarks to understand what sorts of prompts may lead to discriminatory, stereotyping, or demeaning behavior. Depending on the relevance of the exhibited harms, the researcher can then test a number of methods to understand whether these issues can be mitigated or traded off, or if new techniques should be explored or new training datasets created. A benchmark suite can also help the researcher discover fairness-related harms they may not have thought to look for.

Likewise, a non-fairness AI researcher developing a new algorithm or model may use a fairness suite as a tool for monitoring and reporting unintended fairness consequences. At present, a researcher must individually select which benchmarks to test, and this requires specialized knowledge of the fairness landscape. Moreover, the relevant fairness benchmarks may not be obvious at all. For example, seemingly innocuous changes to model structure can create unexpected social disparities,[40] as seen with the "truncation trick" for image quality improvement in generative adversarial networks[41] or even with different computing hardware.[42] Established benchmark suites would make it easier to discover these impacts at the time of development, without having to anticipate which individual benchmarks should be run for which models. Peer reviewers can then choose to judge the research progress of a particular proposed method with this context about the externality effects of that method.

When using benchmark scores, a popular technique is to make sense of trade-offs using visualizations like Pareto frontier curves and radar plots. Visualizations like these can lead to somewhat arbitrary ways of summarizing results that are just as unfounded as averaging, such as the order of each metric on an axis having a large effect.[43] Instead, researchers can work to develop visualizations that help identify which benchmarks are correlated or conflicting, and potentially even incorporate friction by design[44] with respect to quick "top model" selection.

Selecting a model

When selecting a model for development or deployment, rather than defaulting to the results of a leaderboard, practitioners can use benchmark suites to probe a range of potential fairness impacts of a model and make their own contextualized judgments. Benchmark suites can also help to differentiate in the case of model multiplicity (i.e., when multiple models have equal performance under one metric but differ in other metrics due to differences at the individual prediction level).[45,46] Practitioners may often have internal, proprietary benchmarks that can complement public suites. Model cards represent a post-deployment version of this assessment[47] and provide public disclosure of potential fairness issues.

To actually make concrete choices about models by trading off between benchmarks, we echo others' recommendations to perform qualitative analyses of quantitative measurements.[16,48] These qualitatively determined rankings by experts may even compose a specialized leaderboard for a particular application, when the application context and considerations are well understood. Specialized leaderboards should be explicit about which values were prioritized, and may be helpful for filtering and choosing among many possible models. However, leaderboards should not be used without clear and explicit justification because of human cognitive biases that cause us to favor leaderboard results despite any caveats, thus leading to the problems we discussed.

There are a number of different ways that qualitative analyses of quantitative benchmarks can be performed. Transparent deliberations can make clear whose values are prioritized, which can help make space for deferring to epistemic authorities. The epistemic authority of different stakeholders is relevant in terms of their expertise, relationship to the model (e.g., creator, user, auditor), and also their demographic positionality. Given the history of epistemic injustice,[49] practitioners should be prepared to elevate the concerns of historically marginalized and ignored groups that may be the most vulnerable.[49,50] Overarching incentive structures such as profit may come to bear on these, but we should aim to consider collective ways we can continue to elevate social considerations (e.g., consumer power through market backlash), even when they go against what is profitable.[51–56]

TOWARD MORE COMPREHENSIVE BENCHMARK SUITES

With this vision of AI fairness driven by benchmark suites rather than by leaderboards, a remaining question is whether it is possible to build a sufficiently comprehensive benchmark

suite to make this vision a reality. Indeed, developing useful suites is itself an important research problem. A key goal of suite development should be to increase usefulness by expanding (or shrinking) the set of benchmarks included in a suite based on systematic analyses of the coverage, context, and limitations of a suite.[57]

Existing benchmark suites cover multiple facets of what they intend to measure (e.g., trust, capability), but while they aim for comprehensiveness, in practice they are often compiled based on convenience and access. Some, like HELM,[24] do partially map out the negative space of what is not measured. Other recent work takes an important step by displaying a number of dimensions without foregrounding a leaderboard based on score averages.[9]

Building a good benchmark suite is not easy. We argue that there is more to do toward creating suites that increase both coverage and context. This is true even if a benchmark suite is created without a specific deployment setting in mind, although that certainly changes the importance considerations of the various constituent benchmarks. We therefore identify three major shortcomings in current benchmark suites and propose improvements: adopting modality-specific benchmarks, increasing the coverage of harms, and incorporating diverse perspectives.

Benchmarks specific to individual modalities

Fairness evaluations originated in binary classification, and comprehensive collections of fairness metrics often focus only on the binary setting.[58,59] Given the availability of resources for fairness evaluation in the binary setting, evaluations in other settings often repurpose a benchmark measurement that evaluates one modality (e.g., multiple choice classification) for use in evaluating a different modality (e.g., natural language text, image, video).

Unfortunately, these mismatched evaluations create two problems. First, evaluation findings do not transfer well across modalities.[60] For example, existing benchmark suites including Claude 3's model card,[25] TrustLLM,[23] and DecodingTrust,[22] only evaluate fairness in text generation via multiple-choice questions. Claude 3 uses BBQ,[5] the multiple-choice social bias benchmark explained in the introduction, to measure whether an LLM picks the multiple-choice answer that reinforces a stereotype. However, a model's behavior on multiple-choice questions is not necessarily predictive of that model's behavior on natural text output.[60] Additionally, multiple-choice benchmarks fail to capture text-specific linguistic harms (e.g., person-first versus identity-first language when talking about disability[61]).

This is the second problem: modality-specific harms (e.g., those unique to text data) are missed.[62] One prominent example is how measuring fairness in image captioning based on the accuracy of a caption's inferred gender labels[63] misses potential harms of open-ended text generation and other image-specific concerns.[62] Another example is evaluation of robotic systems purely in terms of image and textual metrics[64,65] (e.g., robots instructed to select among physical blocks with pictures of faces from different racial groups for the "criminal" block[66]), which does not evaluate the unique physical harms possible with real robots.

Hence, improving fairness evaluation requires creating benchmarks specific to individual modalities. These evaluations likely require domain-specific insights from outside of computer science. For example, in image generation the analysis of depicted minority group members being either tokenized or genuinely represented could draw from media studies,[67,68] while in text generation, the analysis of stereotypes could leverage theoretical models from social psychology. As an example of operationalizing the latter, the stereotype content model theorizes that all group stereotypes are along the dimensions of warmth and competence (e.g., women are seen to be warm but not competent),[69] and scholars in social psychology have built dictionaries for measuring these concepts in text.[70]

Benchmarks for broader ranges of potential harms

AI fairness research has tended to organize around a few notions of harm. Most early work focused on harms resulting from the differential allocation of resources (e.g., jobs, loans). To expand the focus beyond merely allocational harms, researchers introduced representational harms as those that affect the beliefs, and thus standings, of social groups in society[71] (e.g., search engines overrepresenting White men for "CEO"[72]). Within representational harms, stereotyping has often dominated as the harm of focus.

While this approach has led to valuable research progress, it has also overlooked many other important harms. For example, non-stereotyping representational harms are ignored, like the erasure of social groups (e.g., misrecognizing a hijab as a hoodie in a religious image) or the reification of social groups (e.g., assuming everyone with long hair is a woman).[73]

Additionally, existing benchmarks have often overlooked equity-based harms in favor of equality-based harms.[74] Popular equality-based benchmarks rely on evaluations grounded in invariance and "color-blindness" (i.e., that each demographic group is treated identically). For example, language benchmarks commonly measure changes in model output when the social group of the input is perturbed.[24] Strong benchmarks in other domains like computer vision similarly tend to focus on equality-based assessments.[75] This focus on equal treatment overlooks that certain harms grounded in historical injustice, like misclassifying a person as a gorilla,[76] are more harmful for individuals of certain racial groups than others. Benchmark suites that ignore equity-based harms fail to account for how social context can cause different groups to experience varying degrees and types of harm.

To develop more comprehensive benchmark suites that better capture the full range of potential harms, we advocate for developing benchmarks based on systematic analyses of coverage, rather than merely ad hoc, and convenient, combinations. Proposed taxonomies of harm can help in this goal by mapping out areas of further benchmark development.[73,77,78] Prior work demonstrates a proof-of-concept for building out a benchmark suite based on a taxonomy of representational harms in image captioning.[62] In doing so, it is just as important to record the harms not covered by a benchmark suite as it is to record what is covered. Red-teaming, which has its own set of limitations,[80] can be complementary and help to fill in some holes as well.[79]

In another direction, when thinking about the demographic groups considered, there has been progress in expanding beyond two groups to multiple intersectional ones. However,

as a research community, we are only beginning to think about demographic identities that defy traditional categorization like non-binary gender and residual racial categories like "Multiracial" and "Other."[81,82] These groups may experience specific harms (e.g., misgendering, poor reactions to gender disclosure) that are important to include.[83]

Benchmarks representing different perspectives

Traditional benchmarks usually aggregate ground-truth labels from different human annotators, often by majority vote. However, an annotator's identity (demographic and otherwise) influences which label they will assign.[84–87] For example, stereotypes are culture specific,[88,89] natural language text toxicity can vary based on familiarity and comfort with African American Vernacular English,[90,91] and experiences of stereotype harm can vary based on gender identity.[92] Averaging annotations loses these important distinctions.

In fairness, we should explicitly capture the relevant standpoints that individuals come from by virtue of their lived experiences.[93,94] In instances where combining annotations results in erasing certain voices, more granular benchmark measurements should be considered that disentangle the tension and make it explicit—in other words, by having separate measurements for each side of the tension. In doing so, we will have to be careful about exploitative practices that overburden members of marginalized groups.[95] Instead, we can move toward more substantive participatory approaches by bringing in relevant stakeholders far earlier than simply to annotate.[96,97] This ranges from determining what should be annotated in the first place, to beyond consultation toward inclusion, collaboration, and ownership as is appropriate.[98]

CONCLUSION

Overall, we defend benchmarks but condemn leaderboards in AI fairness. In lieu of leaderboards for making research progress and selecting models, we offer alternative approaches using benchmark suites. This will require improving the coverage and context of benchmark suites by employing modality-specific benchmarks, covering a broader range of harms, and incorporating diverse perspectives. Future work will be needed to assess the missing gaps in benchmark suites, which can motivate the development of even further benchmarks. Benchmark suites are not intended to take the place of ongoing and adaptive system evaluation.[99] Rather, they are complementary sources of information.

In addition to benchmarks, qualitative assessments will remain critical. This is especially true for groups that are very heterogeneous (e.g., non-binary, Disabled, Indigenous) or may contain too few members to perform robust quantitative analyses (e.g., intersectional identities).[100,101]

Taking a broader view than just the model level at which benchmarks often measure, developers should remember to still consider system-level interventions.[102] Systems are broader than models and encompass additional components such as the model interface and how a model is actually used. At the system level, developers can often overcome trade-offs that seemed inevitable at the model level. Let us consider many of the examples we have discussed throughout this paper. In the case of truncation in Generative Adversarial Networks where there is a tension between image quality and diverse racial representation, we could expend resources to collect more images of the underrepresented group and thus improve both image quality and group representation. In the case of image captioning, rather than having a computer vision model either infer gender and risk misgendering someone, or abstain from doing so and thus not provide blind and low-vision users with sufficient context to understand an image fully, the system could prompt self-identification from the depicted individuals and use those identities in any descriptors.[28] As for criminal risk, instead of predicting who might not appear for their court date and preemptively detaining individuals, text reminders of court dates that serve as behavioral nudges can be effective, eliminating much of the need for AI in this use case.[103] We can also imagine even broader solutions than these. All interventions should be done in tandem with structural changes that may even obviate the need for the algorithm in the first place, as well as considerations about where abolitionary acts of refusal toward the technology may be more appropriate.[104]

Changing the status quo usage of leaderboards will be difficult, since leaderboards offer an easy-to-understand, collectively established yardstick for model superiority. Nevertheless, embracing a more nuanced approach to benchmarking—and discarding the idea that a single model can be best suited for all tasks—will better equip researchers and practitioners to confront the complexities of the societal implications of AI. In doing so, we can also hope to avoid algorithmic monocultures where the same model being used across domains may arbitrarily exclude the same individuals,[105,106] and find ourselves with a far greater diversity of models that better reflect society's varied needs.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under grant nos. 1763642, 2112562, and 2145198 to O.R. and a Graduate Research Fellowship to A.W., as well as the Microsoft AI & Society Fellowship to A.W. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. A.H. is an employee of Adobe Research and is affiliated with the University of Washington. We thank Zoya Bylinskii, Vikram Ramaswamy, and Ye Zhu for helpful feedback and comments.

AUTHOR CONTRIBUTIONS

All three authors worked together to brainstorm, ideate, and write the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Benjamin, R. (2019). Race after Technology (Polity).
2. Barocas, S., Hardt, M., and Narayanan, A. (2023). Fairness and Machine Learning: Limitations and Opportunities (MIT Press).
3. Noble, S.U. (2018). Algorithms of oppression: How search engines reinforce racism (NYUPress).
4. Raji, I.D., Bender, E.M., Paullada, A., Denton, R., and Hanna, A. (2021). AI and the everything in the whole wide world benchmark. In Conference on

Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track.

5. Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P.M., and Bowman, S. (2022). BBQ: A hand-built bias benchmark for question answering. In Findings of the Association for Computational Linguistics (ACL), pp. 2086–2105.

6. Hugging Face (2024). Llm safety leaderboard. https://huggingface.co/spaces/AI-Secure/llm-trustworthy-leaderboard.

7. Enkrypt, A.I. (2024). Enkrypt ai safety leaderboard. https://www.enkryptai.com/llm-safety-leaderboard.

8. Zhang, Z., Lei, L., Wu, L., Sun, R., Huang, Y., Long, C., Liu, X., Lei, X., and Tang, J. (2023). andHuang, M. Safetybench. https://llmbench.ai/safety/data.

9. Luxembourg Institute of Science & Technology (2024). LIST LLM leaderboard. https://ai-sandbox.list.lu/llm-leaderboard/.

10. Ding, F., Hardt, M., Miller, J., and Schmidt, L. (2021). Retiring adult: New datasets for fairmachine learning. In Conference on Neural Information Processing Systems (NeurIPS).

11. Bao, M., Zhou, A., Zottola, S., Brubach, B., Desmarais, S., Horowitz, A., Lum, K., and andVenkatasubramanian, S. (2021). It's compaslicated: The messy relationship between rai datasets and algorithmic fairness benchmarks. In Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks.

12. Blodgett, S.L., Lopez, G., Olteanu, A., Sim, R., and Wallach, H. (2021). Stereotypingnorwegian 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, pp. 1004–1015.

13. Alzahrani, N., Alyahya, H.A., Alnumay, Y., Alrashed, S., Alsubaie, S., Almushaykeh, Y., Mirza, F., Alotaibi, N., Altwairesh, N., Alowisheq, A., et al. (2024). When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, pp. 13787–13805.

14. Burnell, R., Schellaert, W., Burden, J., Ullman, T.D., Martinez-Plumed, F., Tenenbaum, J.B., Rutar, D., Cheke, L.G., Sohl-Dickstein, J., Mitchell, M., et al. (2023). Rethink reporting of evaluation results in ai. Science *380*, 136–138.

15. Zhou, K., Zhu, Y., Chen, Z., Chen, W., Zhao, W.X., Chen, X., Lin, Y., Wen, J.-R., and andHan, J. (2023). Don't make your llm an evaluation benchmark cheater. Preprint at arXiv. https://doi.org/10.48550/arXiv.2311.01964.

16. Thomas, R.L., and Uminsky, D. (2022). Reliance on metrics is a fundamental challengefor AI. Patterns *3*, 100476.

17. Ethayarajh, K., and Jurafsky, D. (2020). Utility is in the eye of the user: A critique of nlp leaderboards. In Empirical Methods in Natural Language Processing (EMNLP), pp. 4846–4853.

18. Orr, W., and Kang, E.B. (2024). AI as a sport: On the competitive epistemologies ofbenchmarking. In ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 1875–1884.

19. Berente, N., Kormylo, C., and Rosenkranz, C. (2024). Test-driven ethics for machine learning. Commun. ACM *67*, 45–47.

20. Buolamwini, J., and Gebru, T. (2018). Gender shades: Intersectional accuracy disparitiesin commercial gender classification. In Proceedings of Machine Learning Research, *81*.

21. Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine Bias (Propublica).

22. Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu, C., Xiong, Z., Dutta, R., Schaeffer, R., et al. (2023). DecodingTrust: A comprehensive assessment of trustworthiness in gpt models. In Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track.

23. Sun, L., Yue Huang, H.W., Siyuan Wu, Q.Z., Gao, C., Yixin Huang, W.L., Zhang, Y., Li, X., Liu, Z., Liu, Y., Yijue Wang, Z.Z., et al. (2024). TrustLLM: Trustworthiness in large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.05561.

24. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. (2023). Holistic evaluation of language models. Transactions on Machine Learning Research. Preprint at arXiv *08*.

25. Anthropic (2024). The Claude 3 Model Family: Opus, Sonnet, Haiku (Anthropic). https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

26. Meta (2024). Llama Model Card (Github). https://github.com/meta-llama/llama/blob/main/MODEL_CARD.md.

27. Arthur Team (2024). What's going on with llm leaderboards? Arthur AI. https://www.arthur.ai/blog/whats-going-on-with-llm-leaderboards.

28. Bennett, C.L., Gleason, C., Scheuerman, M.K., Bigham, J.P., Guo, A., and To, A. (2021). "it's complicated": Negotiating accessibility and (mis)representation in image descriptions of race, gender, and disability. In Conference on Human Factors in Computing Systems (CHI), pp. 1–19.

29. Hamidi, F., Scheuerman, M.K., and Branham, S.M. (2018). Gender recognition or genderreductionism? the social implications of automatic gender recognition. In ACM Conference on Human Factors in Computing Systems (CHI), pp. 1–13.

30. Keyes, O. (2018). The misgendering machines: Trans/HCI implications of automatic gender recognition. In Proceedings of the ACM on Human-Computer Interaction, pp. 1–22.

31. Selbst, A.D., danah, boyd, Friedler, S.A., Venkatasubramanian, S., and Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 59–68.

32. Kleinberg, J., Mullainathan, S., and Raghavan, M. (2017). Inherent trade-offs in the fairdetermination of risk scores. In Proceedings of Innovations in Theoretical Computer Science (ITCS).

33. Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data *5*, 153–163.

34. Jacobs, A.Z., and Wallach, H. (2021). Measurement and fairness. In ACM Conference onFairness, Accountability, Transparency (FAccT), pp. 375–385.

35. Mittelstadt, B., Wachter, S., and Russell, C. (2023). The Unfairness of Fair Machine Learning: Levelling Down and Strict Egalitarianism by Default (Michigan Technology Law Review).

36. Goodhart, C. (1984). Problems of monetary management: The uk experience. MonetaryTheory and Practice.

37. Dias Oliva, T., Antonialli, D.M., and Gomes, A. (2021). Fighting hate speech, silencing dragqueens? artificial intelligence in content moderation and risks to lgbtq voices online. Sex. Cult. *25*, 700–732.

38. Rudin, C., and Wagstaff, K.L. (2013). Machine learning for science and society. Mach. Learn. *95*, 1–9.

39. Suresh, H., Movva, R., Dogan, A.L., Bhargava, R., Cruxen, I., Cuba, A.M., Taurino, G., So, W., and D'Ignazio, C. (2022). Towards intersectional feminist and participatory ml: A case study in supporting feminicide counterdata collection. In ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 667–678.

40. Hooker, S. (2021). Moving beyond "algorithmic bias is a data problem". Patterns *2*, 100241.

41. Maluleke, V.H., Thakkar, N., Brooks, T., Weber, E., Darrell, T., Efros, A.A., Kanazawa, A., and Guillory, D. (2022). Studying bias in GANs through the lens of race. In European Conference on Computer Vision (ECCV), pp. 344–360.

42. Nelaturu, S.H., Ravichandran, N.K., Tran, C., Hooker, S., and Fioretto, F. (2023). Onthe fairness impacts of hardware selection in machine learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.2312.03886.

43. to Viz, F.D. (2019). The radar chart and its caveats. From Data to Viz.

44. Cox, A.L., Gould, S.J., Cecchinato, M.E., Iacovides, I., and Renfree, I. (2016). Designfrictions for mindful interactions: The case for microboundaries. In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA), pp. 1389–1397.

45. Breiman, L. (2001). Statistical modeling: The two perspectives. Stat. Sci. *16*, 199–231.

46. Black, E., Raghavan, M., and Barocas, S. (2022). Model multiplicity: Opportunities, concerns, and solutions. In ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 850–863.

47. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Elena Spitzer, I.D.R., and Gebru, T. (2019). Model cards for model reporting. In ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 220–229.

48. Wiggins, C.H. (2018). Ethical Principles, Okrs, and Kpis: What Youtube and Facebook Couldlearn from Tukey (Columbia University Data Science Institute Blog).

49. Fricker, M. (2009). Epistemic Injustice: Power and the Ethics of Knowing (Oxford UniversityPress).

50. Frye, M. (1983). Oppression. The Politics of Reality.

51. Ali, S.J., Christin, A., Smart, A., and Katila, R. (2023). Walking the walk of ai ethics: Organizational challenges and the individualization of risk among ethics entrepreneurs. In ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 217–226.

52. Metcalf, J., Moss, E., and boyd, danah (2019). Owning ethics: Corporate logics, siliconvalley, and the institutionalization of ethics. Soc. Res. *86*, 449–476.

53. Phan, T., Goldenfein, J., Mann, M., and Kuch, D. (2022). Economies of virtue: The circulation of 'ethics' in big tech. Sci. Cult. *31*, 121–135.

54. Belfield, H. (2021). Activism by the AI community: Analysing recent achievements andfuture prospects. In AAAI/ACM Conference on AI, Ethics, and Society (AIES), pp. 15–21.

55. Widder, D.G., Zhen, D., Dabbish, L.A., and Herbsleb, J. (2023). It's about power: Whatethical concerns do software engineers have, and what do they (feel they can) do about them? In ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 467–479.

56. Wang, A., Datta, T., and Dickerson, J.P. (2024). Strategies for increasing corporate responsible ai prioritization. In AAAI/ACM Conference on AI, Ethics, and Society (AIES).

57. Rottger, P., Pernisi, F., Vidgen, B., and Hovy, D. (2024). Safetyprompts: a systematic review of open datasets for evaluating and improving large language model safety. Preprint at arXiv. https://doi.org/10.48550/arXiv.2404.05399.

58. Verma, S., and Rubin, J. (2018). Fairness definitions explained. In FairWare: Proceedings ofthe International Workshop on Software Fairness.

59. Han, X., Chi, J., Chen, Y., Wang, Q., Zhao, H., Zou, N., and Hu, X. (2024). FFB: A fairfairness benchmark for in-processing group fairness methods. In International Conference on Learning Representations (ICLR).

60. Rottger, P., Hofmann, V., Pyatkin, V., Hinck, M., Kirk, H.R., Schütze, H., and Hovy, D. (2024). Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, pp. 15295–15311.

61. Dunn, D.S., and Andrews, E.E. (2015). Person-first and identity-first language: Developing psychologists' cultural competence using disability language. Am. Psychol. *70*, 255–264.

62. Wang, A., Barocas, S., Laird, K., and Wallach, H. (2022). Measuring representationalharms in image captioning. In ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 324–335.

63. Hendricks, L.A., Burns, K., Saenko, K., Darrell, T., and Rohrbach, A. (2018). Women alsosnowboard: Overcoming bias in captioning models. In European Conference on Computer Vision (ECCV), pp. 793–811.

64. Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Advances in Neural Information Processing Systems (NeurIPS), pp. 4356–4364.

65. Kirk, H., Jun, Y., Iqbal, H., Benussi, E., Volpin, F., Dreyer, F.A., Shtedritski, A., and Asano, Y.M. (2021). Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. In Advances in Neural Information Processing Systems (NeurIPS), pp. 2611–2624.

66. Hundt, A., Agnew, W., Zeng, V., Kacianka, S., and Gombolay, M. (2022). Robots enactmalignant stereotypes. In ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 743–756.

67. Wingfield, A.H., and Wingfield, J.H. (2014). When visibility hurts and helps: How intersections of race and gender shape black professional men's experiences with tokenization. Cult. Divers Ethnic Minor. Psychol. *20*, 483–490.

68. Lester, P.M., and Ross, S.D. (2003). Images that Injure: Pictorial Stereotypes in the media (Business & Economics).

69. Fiske, S.T., Cuddy, A.J.C., Glick, P., and Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. J. Pers. Soc. Psychol. *82*, 878–902.

70. Nicolas, G., Bai, X., and Fiske, S.T. (2021). Comprehensive stereotype content dictionariesusing a semi-automated method. Eur. J. Soc. Psychol. *51*, 178–196.

71. Barocas, S., Crawford, K., Shapiro, A., and Wallach, H. (2017). The problem with bias: Allocativeversus representational harms in machine learning. In Proceedings of SIGCIS. Philadelphia, PA.

72. Kay, M., Matuszek, C., and Munson, S.A. (2015). Unequal representation and genderstereotypes in image search results for occupations. In Conference on Human Factors in Computing Systems (CHI), pp. 3819–3828.

73. Katzman, J., Wang, A., Scheuerman, M., Blodgett, S.L., Laird, K., Wallach, H., and Barocas, S. (2023). Taxonomizing and measuring representational harms: A look at image tagging. In AAAI Conference on Artificial Intelligence, *37*, pp. 14277–14285.

74. Watson-Daniels, J. (2024). Algorithmic fairness and color-blind racism: Navigating the intersection. Preprint at arXiv. https://doi.org/10.48550/arXiv.2402.07778.

75. Gustafson, L., Rolland, C., Ravi, N., Duval, Q., Adcock, A., Fu, C.-Y., Hall, M., and Ross, C. (2023). FACET: Fairness in computer vision evaluation benchmark. In International Conference on Computer Vision (ICCV), pp. 20313–20325.

76. Simonite, T. (2018). When It Comes to Gorillas, Google Photos Remains Blind. Wired. https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/.

77. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al. (2021). Ethical and social risks of harm from language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2112.04359.

78. Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla, N., Gallegos, J., Smart, A., Garcia, E., and Virk, G. (2023). Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In AAAI/ACM Conference on AI, Ethics, and Society (AIES), pp. 723–741.

79. Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Ben Mann, E.P., Schiefer, N., Ndousse, K., Jones, A., et al. (2022). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. Preprint at arXiv. https://doi.org/10.48550/arXiv.2209.07858.

80. Friedler, S., Singh, R., Blili-Hamelin, B., Metcalf, J., and Chen, B.J. (2023). Ai Red-Teamingis Not a One-Stop Solution to Ai Harms: Recommendations for Using Red-Teaming for Ai Accountability (Data & Society Policy Brief).

81. Star, S.L., and Bowker, G.C. (2007). Enacting silence: Residual categories as a challengefor ethics, information systems, and communication. Ethics Inf. Technol. *9*, 273–280.

82. Wang, A., Ramaswamy, V.V., and Russakovsky, O. (2022). Towards intersectionality inmachine learning: Including more identities, handling underrepresentation, and performing evaluation. In ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 336–349.

83. Ovalle, A., Goyal, P., Dhamala, J., Jaggers, Z., Chang, K.-W., Galstyan, A., Zemel, R., and Gupta, R. (2023). "i'm fully who i am": Towards centering transgender and non-binary voices to measure biases in open language generation. In ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 1246–1266.

84. Denton, R., D'ıaz, M., Kivlichan, I., Prabhakaran, V., and Rosen, R. (2021). Whose ground truth? accounting for individual and collective identities underlying dataset annotation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2112.04554.

85. D'ıaz, M., Kivlichan, I., Rosen, R., Baker, D., Amironesei, R., Prabhakaran, V., and Denton, R. (2022). Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 2342–2351.

86. Pei, J., and Jurgens, D. (2023). When do annotator demographics matter? measuring theinfluence of annotator demographics with the POPQUORN dataset. In Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII) at ACL, pp. 252–265.

87. Gordon, M.L., Lam, M.S., Park, J.S., Patel, K., Hancock, J.T., Hashimoto, T., and andBernstein, M.S. (2022). Jury learning: Integrating dissenting voices into machine learning models. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI), pp. 1–19.

88. Dev, S., Goyal, J., Tewari, D., Dave, S., and Prabhakaran, V. (2023). Building socioculturally inclusive stereotype resources with community engagement. In Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks, pp. 4365–4381.

89. Sambasivan, N., Arnesen, E., Hutchinson, B., and Prabhakaran, V. (2020). Non-portabilityof algorithmic fairness in india. In Navigating the Broader Impacts of AI Research Workshop at NeurIPS.

90. Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N.A. (2019). The Risk of Racial Bias Inhate Speech Detection (Association for Computational Linguistics (ACL)), pp. 1668–1678.

91. Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., and Smith, N.A. (2022). Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In North American Chapter of the Association for Computational Linguistics (NAACL), pp. 5884–5906.

92. Wang, A., Bai, X., Barocas, S., and Blodgett, S.L. (2023). Measuring machine learningharms from stereotypes: 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).

93. Haraway, D. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. Fem. Stud. *14*, 575–599.

94. Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C.M., Ye, A., Jiang, L., Lu, X., Dziri, N., et al. (2024). A roadmap to pluralistic alignment. In International Conference on Machine Learning (ICML).

95. Jackson, L., Kuhlman, C., Jackson, F., and Fox, P.K. (2019). Including vulnerable populations in the assessment of data from vulnerable populations. Frontiers in Big Data *2*, 19.

96. Sloane, M., Moss, E., Awomolo, O., and Forlano, L. (2022). Participation is not a design fixfor machine learning. In ACM conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO), pp. 1–6.

97. Suresh, H., Tseng, E., Young, M., Gray, M.L., Pierson, E., and Levy, K. (2024). Participation in the age of foundation models. In ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 1609–1621.

98. Delgado, F., Yang, S., Madaio, M., and Yang, Q. (2023). The participatory turn in ai design: Theoretical foundations and the current state of practice. In ACM conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO), pp. 1–23.

99. Jones, E., Hardalupas, M., and Agnew, W. (2024). Under the radar? examining the evaluation of foundation models. Ada Lovelace Institute.

100. Syed, M. (2010). Disciplinarity and methodology in intersectionality theory and research. Am. Psychol. *65*, 61–62.

101. Bowleg, L. (2008). When black + lesbian + woman = black lesbian woman: The methodological challenges of qualitative and quantitative intersectionality research. Sex. Roles *59*, 312–325.

102. Yee, K., Tantipongpipat, U., and Mishra, S. (2021). Image cropping on twitter: Fairnessmetrics, their limitations, and the importance of representation, design, and agency. In ACM Conference on Computer-Supported Cooperative Work And Social Computing (CSCW), pp. 1–24.

103. Fishbane, A., Ouss, A., and Shah, A.K. (2020). Behavioral nudges reduce failure to appear for court. Science *370*, eabb6591.

104. Hampton, L.M. (2021). Black feminist musings on algorithmic oppression. In ACM Conference on Fairness, Accountability, and Transparency (FAccT) (1).

105. Creel, K., and Hellman, D. (2021). The algorithmic leviathan: Arbitrariness, fairness, andopportunity in algorithmic decision making systems. In Conference on Fairness, Accountability, and Transparency (FAccT).

106. Bommasani, R., Creel, K.A., Kumar, A., Jurafsky, D., and Liang, P. (2022). Picking on thesame person: Does algorithmic monoculture lead to outcome homogenization? In Conference on Neural Information Processing Systems (NeurIPS), pp. 3663–3678.

About the authors

Angelina Wang is a postdoctoral fellow at Stanford University. Her research interest is in the area of machine learning fairness and algorithmic bias. She has been recognized by the National Science Foundation Graduate Research Fellowship Program, Electrical Engineering Computer Science Rising Stars, Siebel Scholarship, and the Microsoft AI & Society Fellowship. She earned her Ph.D. in computer science from Princeton University and her B.S. in electrical engineering and computer science from the University of California, Berkeley. She is an incoming assistant professor at Cornell Tech and in the Department of Information Science at Cornell University.

Aaron Hertzmann is a principal scientist at Adobe Research and an affiliate professor at the University of Washington. He received a B.A. in computer science and art and art history from Rice University and a Ph.D. in computer science from New York University. He was a professor at the University of Toronto for 10 years and has also worked at Pixar Animation Studios and Microsoft Research. He has published over 120 papers in computer graphics, artificial intelligence, and art. He is an Association for Computing Machinery Fellow, an Institute of Electrical and Electronics Engineers Fellow, and winner of the SIGGRAPH Computer Graphics Achievement Award.

Olga Russakovsky is an associate professor in the Computer Science Department at Princeton University. She has been awarded the PAMI Young Researcher Award, the National Science Foundation CAREER award, the Anita Borg Emerging Leader Abie Award in Honor of Denice Denton, the Computing Research Association-Widening Participation Anita Borg Early Career Award, the Massachusetts Institute of Technology Technology Review's 35-Under-35 Innovator Award, and the PAMI Everingham Prize. She co-founded and currently serves as board chair of AI4ALL, a nonprofit dedicated to increasing diversity and inclusion in artificial intelligence.