Comments on RFI Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence

Eve Fleisig (UC Berkeley)
Angela Jin (UC Berkeley)
Jessica Dai (UC Berkeley)
Christopher Strong (UC Berkeley)
On behalf of the Automated Decision Systems (ADS) & Society Group at UC Berkeley

Assignment 1. Developing Guidelines, Standards, and Best Practices for AI Safety and Security

Topic 1 - Developing a companion resource to the AI Risk Management Framework

Comment 1.1 - Current techniques and implementations for model validation and verification

Use of simple yet non-comprehensive metrics to measure fairness and bias issues in generative models carry the risk of "fairness washing," in which the lack of evidence that a model causes harm is presented as proof that the model is harmless (Molamohammadi et al., 2023). Because of this, using oversimplified metrics on which models do well carries the risk of allowing model designers to present models as "fair" when they in fact have been completely evaluated. Measures to prevent this issue include (1) using a combination of metrics and evaluation strategies and (2) updating evaluation strategies as flaws are uncovered in older evaluation methods (e.g., Goldfarb-Tarrant et al., 2021; Cheng et al., 2023).

Many neural network verification algorithms were designed in response to the presence of adversarial examples in networks, where small perturbations to the input of networks can yield large and undesirable results (Goodfellow et al., 2014). These kinds of examples have been demonstrated in both digital and physical settings (Eykholt et al., 2018). Much of the work on neural network verification has focused on proving particular input-output properties of networks (Liu et al., 2021) (e.g. verifying there are no adversarial examples within a region). For example, for a network that classifies stop signs within images one might verify that small magnitudes of noise added to these images won't lead to an error. Additional approaches within the field consider similar robustness or stability properties for systems that include learned components and evolve over time. However, there remain significant limitations in this space:

- These frameworks have been designed more around proving robustness than verifying broader concepts of benefit or safety. As a result, for many automated systems it can be difficult to verify the properties that really matter to those impacted by the system. This difficulty can be viewed as stemming from two issues: (i) it can be hard or impossible to select an appropriate set of possible issues to formally check, and (ii) even if you know what you'd like to check you may not be able to represent these kinds of queries in the available frameworks.
- Many of these approaches scale poorly to larger networks.
- Verifying systems that change over time, or that interact with a constantly changing environment, can require strong modeling assumptions or only provide limited guarantees.

As a result, it is unclear the right role for this style of formal verification to guide the design of systems that are broadly safe rather than just robust. By focusing on the impact and use of the system rather than solely on its reliability in the face of error or attacks, perhaps more relevant forms of "verification" can be devised.

References in this subsection:

Cheng, M., Piccardi, T., & Yang, D. (2023). Compost: Characterizing and evaluating caricature in LLM simulations. arXiv preprint arXiv:2310.11501. https://arxiv.org/pdf/2310.11501.pdf.

Goldfarb-Tarrant, S., Marchant, R., Sánchez, R. M., Pandya, M., & Lopez, A. (2020). Intrinsic bias metrics do not correlate with application bias. arXiv preprint arXiv:2012.15859. https://arxiv.org/pdf/2012.15859.pdf.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572. https://arxiv.org/abs/1412.6572.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ... & Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1625-1634). https://openaccess.thecvf.com/content_cvpr_2018/papers/Eykholt_Robust_Physical-World_Attacks_CVPR_2018_paper.pdf.

Liu, C., Arnon, T., Lazarus, C., Strong, C., Barrett, C., & Kochenderfer, M. J. (2021). Algorithms for verifying deep neural networks. Foundations and Trends® in Optimization, 4(3-4), 244-404. https://www.nowpublishers.com/article/Details/OPT-035.

Molamohammadi, M., Taïk, A., Le Roux, N., & Farnadi, G. (2023, October). Unraveling the Interconnected Axes of Heterogeneity in Machine Learning for Democratic and Inclusive Advancements. In Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (pp. 1-12). https://arxiv.org/pdf/2306.10043.pdf.

Comment 1.2 - Content authentication, provenance tracking, and synthetic content labeling

Newer AI-generated text detectors are fairly reliable at detecting AI-generated text that has not been modified (Verma et al., 2023). However, reliable detection of AI-generated text becomes extremely difficult when the AI-generated text is edited by a human (Sadasivan et al., 2023).

Watermarking is an effective strategy for detecting AI-generated text (Aaronson, 2023; Kirchenbauer et al., 2023; Zhao et al., 2023). However, watermarks face issues with robustness, as it is difficult to make watermarks that cannot be removed (Christ et al., 2023). In addition, the benefits of watermarking rely on model designers voluntarily adding watermarks, and it is near-impossible to guarantee that all publicly available models will adhere to this.

References in this subsection:

Aaronson, S. Watermarking of large language models. (2023). Workshop on Large Language Models and Transformers, Simons Institute, UC Berkeley, 2023. URL https://www.youtube.com/watch?v=2Kx9jbSMZqA.

Christ, M., Gunn, S., & Zamir, O. (2023). Undetectable Watermarks for Language Models. *arXiv preprint arXiv:2306.09194*. https://arxiv.org/abs/2306.09194.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I. &amp; Goldstein, T. (2023). A Watermark for Large Language Models. *Proceedings of the 40th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research*. 202:17061-17084 Available from https://proceedings.mlr.press/v202/kirchenbauer23a.html.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*. https://arxiv.org/abs/2303.11156.

Verma, V., Fleisig, E., Tomlin, N., & Klein, D. (2023). Ghostbuster: Detecting Text Ghostwritten by Large Language Models. *arXiv preprint arXiv:2305.15047*. https://arxiv.org/abs/2305.15047.

Zhao, X., Ananth, P., Li, L., & Wang, Y. X. (2023). Provable robust watermarking for ai-generated text. *arXiv preprint arXiv:2306.17439*. https://arxiv.org/pdf/2306.17439.pdf.

Comment 1.3 - Roles that can or should be played by different AI actors for managing risks and harms of generative AI

We emphasize the limitations of self regulation by highlighting several publicly documented instances of retaliation faced by individuals trying to produce reform inside companies (Hao, 2021; Simonite, 2021; Metz, 2021). These instances suggest that companies that develop AI systems cannot be trusted to self-regulate.

A growing body of empirical work studying AI fairness practices at companies that develop AI systems has found that even the best tools for assessing and managing risks and harms of AI systems are ineffective in practice if AI practitioners lack the incentives and resources (e.g., time, financial) to incorporate the tools into their workflows. These findings emphasize the crucial role of organizational- and team-level decisions, culture, and leadership in enabling responsible AI in practice (Vorvoreanu et al., 2023; Madaio et al., 2022; Rakova et al., 2021; Madaio et al., 2020). In light of this work, we suggest NIST consider and distinguish the role of organization-wide and team-level leadership roles, in addition to considering the role of individual AI developers, in guidance on managing risks and harms of generative AI.

References in this subsection:

Hao, K. (2021). She risked everything to expose Facebook. Now she’s telling her story. *MIT Technology Review*. https://www.technologyreview.com/2021/07/29/1030260/facebook-whistleblower-sophie-zhang-global-political-manipulation/.

Madaio, M., Egede, L., Subramonyam, H., Wortman Vaughan, J., & Wallach, H. (2022). Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW1), 1-26. https://dl.acm.org/doi/10.1145/3512899.

Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020, April). 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* (pp. 1-14). https://dl.acm.org/doi/10.1145/3313831.3376445.

Metz, C. (2021). A second Google A.I. researcher says the company fired her. *The New York Times*. https://www.nytimes.com/2021/02/19/technology/google-ethical-artificial-intelligence-team.html.

Rakova, B., Yang, J., Cramer, H., & Chowdhury, R. (2021). 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. https://dl.acm.org/doi/10.1145/3449081.

Simonite, T. (2021). What Really Happened When Google Ousted Timnit Gebru. *Wired*. https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened/.

Vorvoreanu, M., Heger, A., Passi, S., Dhanorkar, S., Kahn, Z., & Wang, R. (2023). Responsible AI Maturity Model. Retrieved from Microsoft website: https://www.microsoft.com/en-us/research/publication/responsible-ai-maturity-model/.

Comment 1.4 - Forms of transparency and documentation for risk management, and best practices to ensure information is shared as needed along the generative AI lifecycle and supply chain

We recommend that NIST -- in addition to drawing on original publications introducing Model Cards, Data Sheets, Data Statements, and more[1] -- also incorporate recent work revising and building on these toolkits through empirical studies with AI practitioners (McMillan-Major et al., 2023) when creating guidelines for transparency and documentation. While many of these tools importantly prompt reflection and ensure accountability from the AI practitioners they are designed for, they often leave community members and advocates out of key questions of risk and harm management. Highlighting this gap, several groups have designed tools to engage these stakeholders in questions of documentation, and, more broadly, questions of reliability, validity, and accountability (Krafft et al., 2021; Shen et al., 2022). We suggest that NIST additionally incorporate these crucial forms of transparency and documentation into its policymaking.

References in this subsection:

Krafft, P. M., Young, M., Katell, M., Lee, J. E., Narayan, S., Epstein, M., ... & Barghouti, B. (2021, March). An action-oriented AI policy toolkit for technology audits by community advocates and activists.

[1] See Table 1 in McMillan-Major et al., 2023 and Appendix A in https://dl.acm.org/doi/pdf/10.1145/3579621 for more comprehensive lists.

In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 772-781). https://dl.acm.org/doi/10.1145/3442188.3445938.

McMillan-Major, A., Bender, E. M., & Friedman, B. (2023). Data Statements: From Technical Concept to Community Practice. *ACM Journal on Responsible Computing*. https://dl.acm.org/doi/10.1145/3594737.

Shen, H., Wang, L., Deng, W. H., Brusse, C., Velgersdijk, R., & Zhu, H. (2022, June). The model card authoring toolkit: Toward community-centered, deliberation-driven AI design. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 440-451). https://dl.acm.org/doi/pdf/10.1145/3531146.3533110.

Topic 2. Creating guidance and benchmarks for evaluating and auditing AI capabilities

Comment 2.1 - Evaluation of the capabilities, limitations, and safety of AI technologies

We recommend that NIST clearly distinguish and define different AI actors. This is especially crucial when considering questions of independence in evaluation and auditing. Measure 1.3 of the NIST AI RMF highlights the importance of participation by independent parties in assessing and monitoring AI risk and related impacts.[2] While we agree with this general need for independent review, we emphasize the need for careful specification of what “independence” requires. For instance, the IEEE 1012-2016 standard for software verification and validation, which has been used to ensure reliability of some of the most complex, safety-critical systems such as nuclear power plants, medical systems and space systems, defines independence by three parameters: technical independence, managerial independence, and financial independence (IEEE Std. 1012-2016). Specifically addressing questions of AI governance, Raji et al. distinguish between “first party”, “second party”, and “third party” audits; and “internal” and “external” oversight (Raji et al., 2022). We recommend that NIST closely engage with these existing definitions when delineating different AI actors and their roles in managing risks and harms of generative AI.

References in this subsection:

Raji, I. D., Xu, P., Honigsberg, C., & Ho, D. (2022, July). Outsider oversight: Designing a third party audit ecosystem for ai governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 557-571). https://arxiv.org/abs/2206.04737.

"IEEE Standard for System, Software, and Hardware Verification and Validation," in *IEEE Std 1012-2016 (Revision of IEEE Std 1012-2012/ Incorporates IEEE Std 1012-2016/Cor1-2017)* , vol., no., pp.1-260, 29 Sept. 2017, doi: 10.1109/IEEESTD.2017.8055462. https://ieeexplore.ieee.org/document/8055462.

[2] “Internal experts who did not serve as front-line developers for the system and/or independent assessors are involved in regular assessments and updates…” (NIST AI RMF, Measure 1.3).

Comment 2.2 - Introduction of biases in AI lifecycle practices and impacts to human and AI teaming performance

The use of human-AI teams carries the risk of automation bias (Goddard, 2012): even if a model exhibits similar biases to a person, people are more likely to take model outputs as neutral and tend not to override model decisions even when they differ from the person's own judgment.

In addition, an algorithm that might appear more accurate than a person under some metric may not necessarily be a better choice. Longstanding research on how humans interact with non-generative models can help to understand how generative models are likely to influence human decisions:

Sentencing algorithms might be more "accurate" than human judges, but humans often account for factors like leniency for first offenses or the potential harms of punishment in ways that algorithms do not (Stevenson et al., 2022). Similarly, work on risk assessment algorithms in the U.S. child welfare system additionally highlight how case workers consider important factors that algorithms predicting risk of child injury do not (Kawakami et al., 2022).

Human overreliance on algorithm outputs is also crucial to consider when understanding impacts of biases on human and AI teaming performance. Prior work has shown that judges using algorithmic risk assessment algorithms changed their sentences in ways that are potentially harmful (Balagopalan et al., 2023). Other work has found overreliance on ML outputs in clinical decision-making, and discusses limitations and considerations surrounding the use of explanations to address overreliance (Jacobs et al., 2021). Generative models present similar concerns, given that there is already research indicating that users put undue trust into model-generated text, including trusting generated misinformation (Spitale et al., 2023).

References in this subsection:

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association: JAMIA, 19(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089.

Balagopalan, A., Madras, D., Yang, D. H., Hadfield-Menell, D., Hadfield, G. K., & Ghassemi, M. (2023). Judging facts, judging norms: Training machine learning models to judge humans requires a modified approach to labeling data. *Science Advances*, *9*(19), eabq0701. https://pubmed.ncbi.nlm.nih.gov/37163590/.

Jacobs, M., Pradier, M. F., McCoy Jr, T. H., Perlis, R. H., Doshi-Velez, F., & Gajos, K. Z. (2021). How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry*, *11*(1), 108. https://www.nature.com/articles/s41398-021-01224-x.

Kawakami, A., Sivaraman, V., Cheng, H. F., Stapleton, L., Cheng, Y., Qing, D., ... & Holstein, K. (2022, April). Improving human-AI partnerships in child welfare: understanding worker practices, challenges, and desires for algorithmic decision support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-18). https://arxiv.org/abs/2204.02310.

Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis) informs us better than humans. *arXiv preprint arXiv:2301.11924*. https://arxiv.org/abs/2301.11924.

Stevenson, M. T., & Doleac, J. L. (2022). Algorithmic risk assessment in the hands of humans. *Available at SSRN 3489440*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3489440.

Comment 2.3 - Impacts on equity

An overlooked impact of generative language models is increased dialect discrimination. Language models exacerbate forms of discrimination not usually considered, against people who speak "low-resource" languages (Hadgu et al., 2023) and against speakers of some varieties of well-resourced languages such as African-American English [AAE] (Deas et al., 2023). Though dialect discrimination often passes unnoticed, it overlaps with well-known axes of discrimination (e.g., discrimination against AAE overlaps with anti-Black racism).

References in this subsection:

Deas, N., Grieser, J., Kleiner, S., Patton, D., Turcan, E., & McKeown, K. (2023). Evaluation of African American Language Bias in Natural Language Generation. *arXiv preprint arXiv:2305.14291*. https://arxiv.org/abs/2305.14291.

Hadgu, A., Azure, P., Gebru, T. (2023). Combating Harmful Hype in Natural Language Processing. In *International Conference on Machine Learning*. PMLR. https://pml4dc.github.io/iclr2023/pdf/PML4DC_ICLR2023_39.pdf.

Comment 2.4 - Considerations for gathering human feedback

Reinforcement Learning from Human Feedback (RLHF) is currently the predominant mechanism for incorporating human feedback into generative language models (Ouyang et al., 2022). However, RLHF has serious flaws: it tends to overgeneralize and necessarily only incorporates feedback from the labelers used (which is often a small, non-representative pool of people) (Casper et al., 2023). Feffer et al. (2024) raise additional concerns regarding AI red-teaming: if not designed thoroughly, it may not be comprehensive enough to prevent serious issues from going unnoticed.

References in this subsection:

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., ... & Hadfield-Menell, D. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*. https://arxiv.org/abs/2307.15217.

Feffer, M., Sinha, A., Lipton, Z. C., & Heidari, H. (2024). Red-Teaming for Generative AI: Silver Bullet or Security Theater?. *arXiv preprint arXiv:2401.15897*. https://arxiv.org/abs/2401.15897.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730-27744. https://arxiv.org/abs/2203.02155.

Comment 2.5 - Optimal composition of AI red teams and teams of human evaluators

There is significant evidence that people have different preferences regarding model behavior based on their personal backgrounds (Goyal et al., 2022; Prabhakaran et al., 2021; Waseem et al., 2016). Because of this, it is critical to ensure teams that provide human feedback or otherwise evaluate language models are stratified across aspects such as their demographic backgrounds. In addition, participatory strategies that allow users and other stakeholders to be involved in the process of model design beyond simply providing feedback help to ensure that voices from a variety of backgrounds are heard (Delgado et al., 2021). Delgado et al. (2023), Feffer et al. (2023), Robertson et al. (2023), and Deng et al. (2023) discuss best practices and challenges in participatory and user-driven techniques. Ensuring that model designers are accountable to feedback is key to ensuring that these approaches result in improved models.

References in this subsection:

Delgado, F., Yang, S., Madaio, M., & Yang, Q. (2021). Stakeholder Participation in AI: Beyond" Add Diverse Stakeholders and Stir". *arXiv preprint arXiv:2111.01122*. https://arxiv.org/abs/2111.01122.

Delgado, F., Yang, S., Madaio, M., & Yang, Q. (2023, October). The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1-23). https://dl.acm.org/doi/10.1145/3617694.3623261.

Deng, W. H., Guo, B., Devrio, A., Shen, H., Eslami, M., & Holstein, K. (2023, April). 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* (pp. 1-18). https://arxiv.org/abs/2210.03709.

Feffer, M., Skirpan, M., Lipton, Z., & Heidari, H. (2023, August). From Preference Elicitation to Participatory ML: A Critical Survey & Guidelines for Future Research. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 38-48). https://dl.acm.org/doi/fullHtml/10.1145/3600211.3604661.

Goyal, N., Kivlichan, I. D., Rosen, R., & Vasserman, L. (2022). Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 1-28. https://arxiv.org/abs/2205.00501.

Prabhakaran, V., Davani, A. M., & Diaz, M. (2021). On releasing annotator-level labels and information in datasets. *arXiv preprint arXiv:2110.05699*. https://arxiv.org/abs/2110.05699.

Waseem, Z. (2016, November). Are you a racist or am I seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science* (pp. 138-142). https://aclanthology.org/W16-5618/.

Robertson, S., Nguyen, T., Hu, C., Albiston, C., Nikzad, A., & Salehi, N. (2023, April). Expressiveness, Cost, and Collectivism: How the Design of Preference Languages Shapes Participation in Algorithmic Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-16). https://dl.acm.org/doi/fullHtml/10.1145/3544548.3580996.

Assignment 2. Reducing the Risk of Synthetic Content

Comment 3.1 - Risks of child sexual abuse, adult content, and discriminatory content in models and datasets

Child sexual abuse and adult content has been found in common image and image+text datasets (Birhane & Prabhu 2020, 2021). We note that language models as well as vision models carry risks of producing such text. Studies have found evidence of these models producing adult content, reproducing conspiracy theories, and perpetuating discrimination against marginalized groups (Cercas Curry & Rieser, 2018; Fleisig et al., 2023; Gehman et al., 2020). Though model audits are critical; we also note that the *datasets* on which models are trained are often also unavailable to the public. Given that many issues in downstream models stem from issues in their training data, allowing auditors to have access to the training data, not just the trained model, is crucial.

References in this subsection:

Birhane, A., Prabhu, V. U., & Kahembwe, E. (2021). Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*. https://arxiv.org/pdf/2110.01963.pdf.

Prabhu, V. U., & Birhane, A. (2020). Large image datasets: A pyrrhic win for computer vision?. *arXiv preprint arXiv:2006.16923*. https://ieeexplore.ieee.org/document/9423393.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*. https://aclanthology.org/2020.findings-emnlp.301/.

Fleisig, E., Amstutz, A., Atalla, C., Blodgett, S. L., Daumé III, H., Olteanu, A., ... & Wallach, H. (2023). Fair-Prism: Evaluating fairness-related harms in text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics*. https://aclanthology.org/2023.acl-long.343.pdf.

Curry, A. C., & Rieser, V. (2018, June). # MeToo Alexa: how conversational systems respond to sexual harassment. In *Proceedings of the second acl workshop on ethics in natural language processing* (pp. 7-14). https://aclanthology.org/W18-0802/.

Assignment 3. Advance Responsible Global Technical Standards for AI Development

Comment 4.1 - Best practices regarding data capture

Existing practices for collecting and labeling data for use in generative models carry serious implications for labor protections. Workers who perform data labeling (often remotely, as crowdworkers) are often in a precarious position: wages are very low (often below minimum wage), labor protection is nonexistent, and the content they are asked to label can border on traumatic (Xia et al., 2017; Huang et al., 2023; Perrigo, 2022).

References in this subsection:

Xia, H., Wang, Y., Huang, Y., & Shah, A. (2017). " Our Privacy Needs to be Protected at All Costs" Crowd Workers' Privacy Experiences on Amazon Mechanical Turk. *Proceedings of the ACM on human-computer interaction*, *1*(CSCW), 1-22. https://dl.acm.org/doi/pdf/10.1145/3134748.

Huang, O., Fleisig, E., & Klein, D. (2023). Incorporating Worker Perspectives into MTurk Annotation Practices for NLP. *arXiv preprint arXiv:2311.02802*. https://arxiv.org/abs/2311.02802.

Perrigo, B. (2022, February 14). Inside Facebook's African Sweatshop. Time. https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/.

Comment 4.2 - Application-specific standards: language models

Choosing tools designed to filter out harmful content from generative language models requires caution, as many popular older tools (e.g. Jigsaw/Perspective) are known to have serious issues with classifying AAE as hate speech yet continue to be used. Use newer models that have implemented mitigations for this issue.

References in this subsection:

Schlesinger, A., O'Hara, K. P., & Taylor, A. S. (2018, April). Let's talk about race: Identity, chatbots, and AI. In Proceedings of the 2018 chi conference on human factors in computing systems (pp. 1-14). https://dl.acm.org/doi/10.1145/3173574.3173889.

Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. arXiv preprint arXiv:1905.12516. https://aclanthology.org/W19-3504/.

Thiago, D. O., Marcelo, A. D., & Gomes, A. (2021). Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. Sexuality & culture, 25(2), 700-732. https://www.researchgate.net/publication/345501707_Fighting_Hate_Speech_Silencing_Drag_Queens_Artificial_Intelligence_in_Content_Moderation_and_Risks_to_LGBTQ_Voices_Online.

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., ... & Gardner, M. (2021). Documenting large webtext corpora: A case study on the colossal clean crawled corpus. arXiv preprint arXiv:2104.08758. (also talks about issues with datasets). https://arxiv.org/abs/2104.08758.