Sociotechnical and System Considerations for End-to-end Fairness: An Annotated Reading List

AUDREY CHANG, Harvard University
CONAN LU, Harvard University
GUSTAVO MERINO MARTINEZ, Harvard University
JADE NAIR, Harvard University
EMMANUEL RASSOU, Harvard University
LIA ZHENG, Harvard University

The following readings shed light on key challenges and opportunities in advancing AI ethics and transparency in implementation. Mainly centered around algorithmic bias, these readings encompass methods, case studies, and frameworks to develop public awareness of AI harms, public participation in model development, developer awareness of downstream harms, and regulatory standards for algorithmic integration into real-world systems. The papers span considerations at various steps in the ML development cycle, from pre-development harm identification, model training, to post-deployment sociotechncial implementations. Readings were selected as a part of a reading group run by Harvard's Responsible Computing Collective, ReCompute.

Additional Key Words and Phrases: Algorithmic Fairness, Public Participation, Model Development Considerations, Regulatory Standards

[1] E. Corbett, E. Denton, and S. Erete, "Power and public participation in ai," in *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, ser. EAAMO '23. New York, NY, USA: Association for Computing Machinery, 2023.

The reading applies Arnstein's Ladder of Citizen Participation by surveying 21 papers and categorizing them into each rung. It discusses the utility of machine learning models from a sociological lens, along with calling for more social participation in the development and implementation of artificial intelligence engines/research. Rungs range from "manipulation" to "citizen control," in which citizens have influence over the research question. The central idea is that participation is only genuine if it involves the redistribution of power. The reading uses the rhetoric of designing a system for a group rather than with, and informing a group rather than collaborating with. Our reaction is that the decision to move "up" the ladder is very context-dependent. Since the reading grouped most papers in the middle rungs, but those are also labeled as tokenization, it suggests this is the current status quo. Alternatively, we think that there should be a second dimension which considers the type of application. For example, translation applications should include more perspectives, whereas something like autonomous driving models would likely respond negatively. Even though the practical limitations of such research is overlooked, the potential benefits of instating public transparency as the norm in training models democratically has potential to create a less polarized society, thereby redistributing the power among those who affect AI development.

[2] M. Feffer, N. Martelaro, and H. Heidari, "The ai incident database as an educational tool to raise awareness of ai harms: A classroom exploration of efficacy, limitations, & future improvements," in *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 2023, pp. 1–11.

This presents an intriguing hypothetical regarding the implications of unethical AI practices and the consequences of leaving MLMs unchecked. The reading explores AI Incident Database, AIID, as an educational tool to raise awareness about potential AI harms and facilitating discourse on AI safety and accountability. To do so, they test the merits of the AIID as a tool for a specific subgroup of AI-interested students. The AIID is regulated in that it reviews incoming reported incidents. However, anyone can self-report harms via social media, bringing up the question of whether users are well-equipped to notice when their experienced harms are caused by AI, and to argue its direct impact. Overall, the AIID curates important lessons and considerations to be aware of for the future AI design. This experiment allows us to ask how we can better integrate AI ethics into CS courses relevantly

Authors' addresses: Audrey Chang, audreychang@college.harvard.edu, Harvard University; Conan Lu, conanlu@college.harvard.edu, Harvard University; Gustavo Merino Martinez, gmerinomartinez@college.harvard.edu, Harvard University; Jade Nair, jnair@college.harvard.edu, Harvard University; Emmanuel Rassou, emmanuel_rassou@college.harvard.edu, Harvard University; Lia Zheng, liazheng@college.harvard.edu, Harvard University.

(i.e., how effective is Harvard's Embedded EthiCS?) The implications of this piece reach far beyond the intended scope of developers programmers; a universal AI ethics curriculum could have a grand impact on the future of policymaking and machine learning overall.

[3] N. Pagan, J. Baumann, E. Elokda, G. De Pasquale, S. Bolognani, and A. Hannák, "A classification of feedback loops and their relation to biases in automated decision-making systems," in *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 2023, pp. 1–14.

The reading establishes a thorough and well-drawn outline of various machine learning models, defining categorical terms for feedback loops as well as potential biases they may fall victim to. It uses recommender systems (RS) on online content platforms as a case study for the effect of feedback loops on bias. For example, feedback loops can perpetuate bias if the results of the same data are re-used, but can actually help to mitigate it in cases where new, less-biased training data is introduced. The reading identifies representation bias, historical bias, and measurement bias as the three possible downstream effects of various types of feedback loops. Although it explains the relevance of each feedback loop with a real life scenario of it in play, it does not dive into strategies for mitigating these loops once they are correctly classified. A key inquiry raised is how to balance the need for fair decision-making with the goal of accuracy in a model, as fixing feedback loops may inherently tradeoff with accuracy.

[4] A. Wang, X. Bai, S. Barocas, and S. L. Blodgett, "Measuring machine learning harms from stereotypes: requires understanding who is being harmed by which errors in what ways," *arXiv preprint arXiv:2402.04420*, 2024.

Through a series of survey studies, the paper identifies what types of stereotyping is harmful and why, characterizing their impacts as prescriptive, proscriptive, or perpetuating of a negative trait. The text introduces a framework that distinguishes between pragmatic and experiential harm and describes the limits of designing exclusively to reduce experiential harm, citing instances where individuals react negatively to non-normative images. Results promote discussion of how to decide which stereotypes are acceptable for AI systems to propagate. Furthermore, it highlights consequences posed by machine learning errors, considerations when mitigating algorithmic bias, and proposes measuring their impact through both quantitative and qualitative metrics.

[5] M. Zilka, C. Ashurst, L. Chambers, E. P. Goodmann, P. Ugwudike, and M. Oswald, "Exploring police perspectives on algorithmic transparency: A qualitative analysis of police interviews in the uk," in *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, ser. EAAMO '23. New York, NY, USA: Association for Computing Machinery, 2023.

The study explores the implications of police adoption of the UK Government's 'Algorithmic Transparency Recording Standard,' which aims to standardize government communication about algorithmic tools to the public. Through semi-structured interviews with stakeholders from UK policing and related commercial entities, the research identifies the Standard's potential benefits, risks, and challenges for law enforcement, as well as areas for improvement. It notes police reluctance to increase transparency, possibly rooted in concerns about administrative burdens and misunderstood biases in their models. It proposes specific updates to the Standard, advocating for enhanced user guidance and clarification of the Standard's scope, providing exemptions for sensitive contexts and additional fields for data and model comprehensiveness, ensuring resource availability for compliance, and addressing supplier responsibilities in procurement contracts. However, the study acknowledges limitations such as sample bias and the qualitative nature of sourcing interviewees, resulting in a detailed framework for transparency in UK law enforcement with limited substantiation due to the study's narrow scope. The text motivates important questions to the extent that transparency should be required: these models have huge influence on people's lives, so does there need to be a threshold for interpretability in models, and is that even tractable for public use?