Re-imagining Fairness in Machine Learning: A Framework for Building in Socio-cultural and Contextual Awareness

COREY JACKSON, University of Wisconsin - Madison, USA
TALLAL AHMED, University of Wisconsin - Madison, USA
DEVANSH SAXENA, Carnegie Mellon University, USA

Machine learning algorithms have become a central component of decision-making, permeating many facets of life, deciding who people date, their creditworthiness, etc. Initially heralded as an innovation in reducing human cognitive and social bias in decision-making, the data used to construct such models and the models themselves are often embedded with the same biases the models had proposed to solve. Developers have looked toward implementing mathematical models to address algorithmic bias. Fairness is an amorphous and contextually defined term dependent on socio-historical and other contexts. We propose a system in which developers and the public collaborate to build a set of machine learning norms around data, models, and interfaces that are socially, historically, contextually, and geographically aware. The research outlined in this proposal is the first step toward building such a system. This position paper describes a plan to understand fairness as a socially, geographically, and contextually situated construct. Our research focuses on building knowledge about fairness to supplement existing approaches that enhance fairness in ML auditing processes.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; **User studies**; • **Computing methodologies → Machine learning**.

Additional Key Words and Phrases: Fairness, Global South, AI-Audits, Machine Learning

ACM Reference Format:
Corey Jackson, Tallal Ahmed, and Devansh Saxena. 2018. Re-imagining Fairness in Machine Learning: A Framework for Building in Socio-cultural and Contextual Awareness. In . ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

While the use of AI systems in decision-making contexts has grown considerably, mounting evidence suggests that decisions informed in part or wholly by AI systems can lead to disparate risks and harms to individuals and groups. ML developers have recently turned to AI audits to evaluate their algorithms to mitigate such harms. It is well known that when biased datasets are used to make predictions, AI systems can embed human and societal biases that often result in decisions that are "unfair" or "unjust" [1, 7]. The COMPAS software, for example, uses predictive modeling to assess recidivism risks. Its implementation has been criticized because earlier versions of the model tended to overestimate the recidivism risks of black individuals compared to white individuals [1]. In the United Kingdom, the Ofqual grading algorithm relied on historical grade distributions, which deflated grades for state schools and inflated grades for private school students – leading to unfair treatment of students of lower socioeconomic backgrounds who most often attend state schools [4].

Recent research highlights the sources of "unjust" outcomes and often points to issues such as existing biases in data used for training models. Data used in training AI systems often reference sensitive classes of data about individuals, which, historically, have been used to discriminate against individuals and groups. In the United States, discrimination based on characteristics such as race and gender has long been a source of unequal treatment for many individuals and groups. Evaluating models before implementation has gained traction in developer communities. ML developers often evaluate their algorithms against fairness metrics (fairness interventions or constraints), seeking to quantify aspects of individual and group fairness before implementation. As a form of auditing, fairness evaluations have successfully reduced bias and made fairer decisions. While fairness audits have been beneficial, we see *at least* two challenges to ensuring that all AI systems remain fair.

First, *no universal definition of fairness applies to all contexts*. Fairness is often subjective, and what is considered fair can vary greatly depending on cultural, social, and individual perspectives. Current fairness interventions tend to reduce the issue of fairness to mathematical and statistical techniques, ignoring the historical, social, and structural factors that occasion biases. ML developers often need to rely on their understanding of the context to select which fairness metrics to evaluate their algorithm against. While optimizing around the statistical definition of fairness, the sensitive features to be optimized may be incompatible with what a population regards as fair. Context also plays a crucial role in shaping human perspectives on fairness. In most instances, ML developers are removed from the socio-historical context in which training data are collected. Ignoring these important antecedents of unfairness ensures most models include unjust or biased outcomes.

Second, *much of the current theoretical and empirical work has been grounded in US and Euro-centric views of social and legal conceptualizations of fairness*. Fairness, however, varies among individuals from different socioeconomic, cultural, or ethnic backgrounds. As AI systems span beyond cultural boundaries, fairness interventions may still not solve the problem as cultures worldwide offer different lenses of fairness. US-Euro-centric definitions may not encapsulate the issue such that technical solutions would sufficiently solve the problem. For example, many fairness interventions seek adjustments to sensitive attributes referenced in legislation such as the United States Equal Credit Opportunity Act, which makes discrimination unlawful concerning any aspect of a credit application based on race, color, religion, national origin, sex, marital status, and age as evaluative criteria. Empirical research on fairness has demonstrated that people in different places and cultural backgrounds may emphasize different facets of fairness. Leung and Stephan [5] argued that pluralistic societies in the East differ from the egalitarian cultures of the West; therefore, different forms of justice, distributive, procedural, and retributive, are conceptualized and achieved differently. Blake et al. compared the acquisition of fairness behavior across seven different cultures and found that advantageous inequity aversion was more prevalent in Western cultures than in Eastern cultures. Therefore, we expect cultural variations and national boundaries to mediate the importance placed on different facets of fairness. Other facets of fairness that may have different salience include sanctity and reconciliation [6, 8]. Even when empirical research into applying individual and group fairness metrics is available, their applications may not capture moral concerns intrinsic to fairness. Some concerns may be more important or valued differently across contexts and cultures (especially when the context and cultures differ from the one in which they are developed or from which training data are drawn). Given these limitations, the field needs to (1) better align fairness metrics with the context in which AI systems are implemented and (2) ensure bias reduction strategies are inclusive of multiple views of fairness.

2 CROWDSOURCING FAIRNESS AUDITS

To address the above-mentioned challenges, we propose crowdsourcing attitudes about fairness in different decision-making contexts and with diverse populations. We view fairness as contextually situated and attitudes towards fairness as dependent on socio-historical and other contexts. We are researching how fairness attitudes differ in context and geography. We propose to build a system in which model builders and the public collaborate to build a set of socially, historically, contextually, and geographically situated norms of fairness. Augmenting the standard machine learning workflow, our research focuses on revising the machine learning workflow to include crowdsourced attitudes about fairness to supplement existing approaches that enhance fairness in ML auditing processes.

The research draws inspiration from the Moral Machine [2], an online platform for gaining human perspectives on moral dilemmas through human judgments of the trolley problem for self-driving cars. Users are presented with scenarios and asked to judge what an autonomous vehicle should decide—in 1, choosing to run over five pedestrians, resulting in their death, or crashing the vehicle into a barrier, resulting in the death of the vehicle's passengers. These judgments help designers of AI systems gather public opinions on how self-driving cars should behave. Moral Machine collects millions of judgments from a global population to gain insights into these differences and similarities in ethical preferences. In a study of Moral Machine judgments, Awad et al. [2] identified strong preferences among respondents for saving human lives, saving more lives, and saving young lives. The study also highlighted preferences based on demographics, culture, and geography, suggesting that achieving consensual machine ethics is feasible.

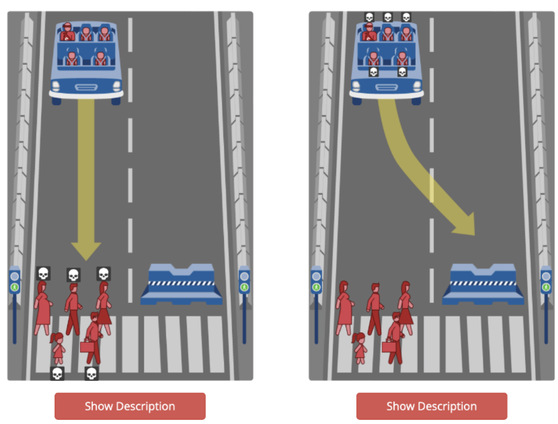

Fig. 1. The Moral Machine judgment interface.

Similarly, we envision a system that builds and catalogs knowledge about fairness across various decision-making contexts and cultures. We are planning a research study to evaluate the feasibility of crowdsourcing fairness through a process similar to Moral Machine. In our system, users will be presented with scenarios where they will be asked to decide on preferable outcomes for the model. For example, in a healthcare AI system, we might ask, "You're a health insurance executive reviewing an algorithm determining premium rates. The criteria for evaluating patients include age, health conditions, and lifestyle factors. Which of the options below best aligns with how you would design the algorithm?" The response options each map to a common fairness metric (e.g., disparate impact, statistical parity, equal opportunity, and equal odds) but do not explicitly reference fairness or the mathematical model. In the above example, the response option mapping to disparate impact would read, "Ensuring individuals from each age group have proportional premium rates." These scenarios and responses would be aggregated to identify which metric most aligns with their attitudes toward fairness. In addition to collecting fairness judgments for each scenario, we intend to collect demographic information (e.g., gender, birth country) to evaluate fairness attitudes across cultures.

We envision this project to follow a research-through-design (RtD) process [9] where we iteratively improve how we formulate our research questions (RQs) and scenarios by integrating design prototypes. For this project, we will first create low-fidelity prototypes and hold a series of workshops with domestic students and international students at the University of Wisconsin-Madison and gather their feedback. We will recruit students from diverse backgrounds, including those whose home countries are not considered Western nations. This approach will allow us to observe students' interaction with the prototype, assess if the process advances smoothly, and whether students can freely ideate their perceptions of fairness in the given scenarios. There are several RtD questions that we are considering as we approach this project, but we recognize that we will be reframing some of them after conducting the initial workshops [10]. We list some of these RQs regarding the design workshops -

(1) Are there key differences between domestic and international students' fairness perceptions? What are the underlying values that motivate these differences?
(2) Should we consider providing our participants with a brief overview of fairness definitions before starting the workshops?
(3) Will that lead to more nuanced conversations about fairness, or would we be priming our participants to think about fairness in set ways?

To answer the questions above, we have considered parallel prototyping [3]; however, in prior projects, we have found that the prototypes can result in two distinct sets of information based on the participants' socio-cultural background. How do we effectively account for these divergences when undertaking systems design? Answers to these questions will help us better understand how participants from different socio-cultural backgrounds perceive fairness concerns and further allow us to create more nuanced scenarios (similar to the Morale Machine) where there would be ambiguity about the 'right answer' and the participants must resolve this ambiguity by employing their core values to arrive at what they believe to be the fair answer.

The research outlined above and our participation in the workshop will address a growing societal concern: the pernicious effects of AI systems and their ability to exacerbate biases that negatively impact marginalized groups. As AI-based decision-making systems are developed, systematically studying and evaluating systems is necessary. Furthermore, research is needed that centers on the concerns of individuals and groups not involved in constructing AI systems but are impacted by them.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. (2016).
[2] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature* 563, 7729 (2018), 59–64.
[3] Steven P Dow, Alana Glassco, Jonathan Kass, Melissa Schwarz, Daniel L Schwartz, and Scott R Klemmer. 2010. Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *ACM Transactions on Computer-Human Interaction (TOCHI)* 17, 4 (2010), 1–24.
[4] Upol Ehsan, Ranjit Singh, Jacob Metcalf, and Mark Riedl. 2022. The algorithmic imprint. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1305–1317.
[5] Kwok Leung and Walter G Stephan. 2001. Social justice from a cultural perspective. *The handbook of culture and psychology* (2001), 375–410.
[6] Donald J Lund, Lisa K Scheer, and Irina V Kozlenkova. 2013. Culture's impact on the importance of fairness in interorganizational relationships. *Journal of International Marketing* 21, 4 (2013), 21–43.
[7] Cathy O'neil. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.
[8] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 315–328.
[9] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. 2007. Research through design as a method for interaction design research in HCI. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 493–502.

[10] John Zimmerman, Aaron Steinfeld, Anthony Tomasic, and Oscar J. Romero. 2022. Recentering Reframing as an RtD Contribution: The Case of Pivoting from Accessible Web Tables to a Conversational Internet. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–14.

Received 15 September 2023