VIEWPOINT

Participatory Machine Learning Using Community-Based System Dynamics

VINODKUMAR PRABHAKARAN AND DONALD MARTIN JR.

The pervasive digitization of health data, aided with advancements in machine learning (ML) techniques, has triggered an exponential growth in the research and development of ML applications in health, especially in areas such as drug discovery, clinical diagnosis, and public health.[1] A growing body of research has shown evidence that ML techniques, if unchecked, have the potential to propagate and amplify existing forms of discrimination in society, which may undermine people's human rights to health and to be free from discrimination.[2] We argue for a participatory approach that will enable ML-based interventions to address these risks early in the process and to safeguard the rights of the communities they will affect.

The promise of machine learning is its ability to efficiently comb through data to find valuable patterns and insights that the machine may use to make predictions or to aid humans in making decisions. However, data reflect numerous societal and human biases that shape their generation, availability, collection, synthesis, and analysis. Machines learn insights based on correlations in data; but most current ML algorithms do not have a means to distinguish between correlations that are mere reflections of these societal biases (for example, racial and gender disparities in society) and those that are causal and reliable insights on which to base their decisions. This is especially problematic in high-stakes domains such as health, where propagating and amplifying such societal biases may disproportionately harm those who are already facing discrimination in society.

For instance, a recent study found that an ML-based health care risk-assessment tool used in the United States exhibited racial bias against Black Americans, denying them access to special programs and resources.[3] The goal of the risk-assessment tool was to improve care for patients with complex health needs while reducing overall costs by connecting high-risk patients with special programs and resources. During the ML problem formulation, this strategic goal was reduced to identifying patients who had the highest health care costs, relying on the implicit causal theory held by the developers that patients with more complex health needs would have spent more on health care in the past. However, this inference failed to consider the historic disparities in health care access (among other things) that Black individuals face in the US health care system and the dynamically complex ways that such disparities affect their spending on health care.

Vinodkumar Prabhakaran is a Research Scientist at Google Research, San Francisco, USA.

Donald Martin Jr. is a Technical Program Manager at Google Trust and Safety, San Francisco, USA.

Please address correspondence to Vinodkumar Prabhakaran. Email: vinodkpg@google.com.

Competing interests: None declared.

Consequently, the algorithm tended to mistakenly construe Black individuals as not being high-risk patients, further denying them access to special programs and resources.

One mistake made in the design of the aforementioned tool was the decision to use health care costs (spending) as a proxy variable for health care needs. Relying on such simplified models of the societal context in which ML-based interventions will be deployed fails to account for the dynamically complex nature of society and the various factors affecting how health care needs may be reflected in data. More importantly, the assumptions that guided this choice emerged from an opaque and iterative process among key internal stakeholders (often limited to product managers, business analysts, computer scientists, and ML practitioners) resulting from their cumulative lived experiences and reflecting their world views and biases. These stakeholders often lack the subject-matter expertise or lived experiences required to comprehensively approximate and account for the various peripheral stakeholders whom their interventions will affect, especially the communities that are already subject to social discrimination. For instance, if the problem-formulation step had facilitated the equitable participation of diverse communities with lived experiences within the US health care system, the developers could have flagged that the above assumption regarding health care spending was flawed. This capability gap is a core issue that contributes to the recurring blind spots of tech interventions in society.

Toward participatory methods in machine-learning fairness

Fairness failures in deployed ML systems may have negative impacts on human rights, such as the right to the highest attainable standard of health and the right to be free from discrimination. However, explicit human rights considerations have largely been absent within the community of technical researchers working to ensure fairness in machine learning. Computer scientists often focus on the biases in their models and attempt to mitigate them through algorithmic means. Such purely observational and statistical approaches may be inadequate when considering normative, constitutive, process-oriented, and socially constructed concepts such as fairness and equity.

As a solution, we recently proposed a complex adaptive system (CAS)-based model of societal context.[4] CASes are *complex* in the sense that they are made up of components that are directly or indirectly related in a causal network, and the behavior of the system cannot be predicted based solely on the behavior of its components; and they are *adaptive* in the sense that they adapt to the changes in their environment by mutating or self-organizing their internal structures. The CAS-based model has been successfully applied to model social systems of varying sizes and complexity, from individual organizations to large health care systems. A key component in our model is the causal theories that human agents hold about the cause-to-effect relationships between various factors that cause or lead to specific problems in society. In the health care example above, the assumption that more complex health needs would lead to increased health spending for all groups is analogous to a causal theory in the CAS model. Highlighting causal theories as a key component of societal context emphasizes the importance of more complete causal theories and incentivizes making them explicit for scrutiny, critique, and improvement.

In order to mitigate the negative consequences of incomplete causal theories described above, we recently proposed community-based system dynamics (CBSD) as a practice that could supply diverse sources of causal theories to core decision-making steps during the ML development process.[5] CBSD is a participatory method that relies on group modeling sessions involving diverse stakeholders, with the goal of developing a shared understanding of a *complex adaptive* problem by making the causal theories held by participants explicit. It relies both on informal maps and diagrams to make everyone's causal theories more explicit and on formal models with computer simulation to uncover the dynamics of complex problems from a feedback perspective. To uncover and understand

feedback processes, CBSD uses a series of graphical tools with varying degrees of formalism, requiring modelers to make their causal theories explicit. This reliance on visual diagramming in CBSD emphasizes transparency and facilitates the engagement of diverse stakeholders to add, revise, and critique causal theories. Moreover, a strength that this methodology shares with other causal modeling approaches is the correspondence between its visualizations and their underlying mathematical representations, which allows stakeholders to develop deep insights about important data for collection and consideration and to simulate the impact of their interventions.

Rather than merely gathering insights through participation, CBSD seeks to co-create solutions in a way that ensures communities' active involvement. The process often results in counterintuitive insights about the problem space and has successfully led to solutions that challenge conventional wisdom in numerous interventions in public health and social work.[6] Employing CBSD to build fairer technologies means that stakeholders get to define and negotiate together what fairness means in the contexts where these technologies are applied. For instance, an initial CBSD-oriented workshop was held at the Data for Black Lives II conference in January 2019. Attended by about 70 participants, this workshop facilitated group model-building exercises on the topic of the racial wealth gap in the United States. A subset of conference participants continued on for a months-long CBSD effort on the topic of racial bias in artificial intelligence and its implications for health disparities. The modeling process and outcomes were then presented at the 2020 Conference of the System Dynamics Society.[7] This work demonstrated how employing CBSD to center the discussion of data and health care on people and their experiences helped derive important structural insights into how ML-based interventions in health care may perpetuate or exacerbate racial biases. Specifically, the CBSD process identified collective memory of community trauma (through deaths attributed to poor health care) and negative experiences with health care as endogenous drivers of seeking treatment and experiencing effective care, which in turn affect the availability and quality of data for algorithms.

We believe that a proactive, participatory, rights-based approach to ML fairness will provide the much-needed grounding for a set of globally salient and cross-culturally accepted values and principles and will help orient the conversation toward humans and the risks to their rights rather than machines and the risks of their biases. Businesses have the responsibility to protect and respect human rights, as outlined in the United Nations Guiding Principles on Business and Human Rights.[8] Effective and scalable participatory methods such as CBSD may help bring forth the perspectives of marginalized communities during the earliest stages of the product development process, enabling the co-creation of solutions by technologists and communities. These efforts could inform companies' approaches to evaluating potential human rights impacts across the product life cycle.

Disclaimer

Any opinions, findings, conclusions, or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of their employer.

References

1. F. Wang and A. Preininger, "AI in health: State of the art, challenges, and future directions," *Yearbook of Medical Informatics* 28/1 (2019), p. 16.

2. F. A. Raso, H. Hilligoss, V. Krishnamurthy, et al., "Artificial intelligence and human rights: Opportunities and risks," Berkman Klein Center Research Publication 2018-6 (2018).

3. Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science* 366/6464 (2019), pp. 447–453.

4. D. Martin Jr., V. Prabhakaran, J. Kuhlberg, et al., "Extending the machine learning abstraction boundary: A Complex systems approach to incorporate societal context," arXiv:2006.09663 (2020).

5. D. Martin Jr., V. Prabhakaran, J. Kuhlberg, et al., "Participatory problem formulation for fairer machine learning through community based system dynamics," ICLR Work-

shop on Machine Learning in Real Life, April 26, 2020.

6. J. B. Homer and G. B. Hirsch, "System dynamics modeling for public health: Background and opportunities," *American Journal of Public Health* 96/3 (2006), pp. 452–458.

7. J. Kuhlberg with I. Headen, E. Ballard, and D. Martin Jr., "Advancing community engaged approaches to identifying structural drivers of racial bias in health diagnostic algorithms," paper presented at the 2020 Conference of the System Dynamics Society, July 19–24, 2020. Available at https://proceedings.systemdynamics.org/2020.

8. Office of the United Nations High Commissioner for Human Rights, *Guiding principles on business and human rights* (New York: United Nations, 2011).