Human Comprehension of Fairness in Machine Learning

Debjani Saha
University of Maryland
dsaha@cs.umd.edu

Candice Schumann
University of Maryland
schumann@cs.umd.edu

Duncan C. McElfresh
University of Maryland
dmcelfre@math.umd.edu

John P. Dickerson
University of Maryland
john@cs.umd.edu

Michelle L. Mazurek
University of Maryland
mmazurek@cs.umd.edu

Michael Carl Tschantz
ICSI
mct@icsi.berkeley.edu

ABSTRACT

Bias in machine learning has manifested injustice in several areas, with notable examples including gender bias in job-related ads [4], racial bias in evaluating names on resumes [3], and racial bias in predicting criminal recidivism [1]. In response, research into algorithmic fairness has grown in both importance and volume over the past few years. Different metrics and approaches to algorithmic fairness have been proposed, many of which are based on prior legal and philosophical concepts [2]. The rapid expansion of this field makes it difficult for professionals to keep up, let alone the general public. Furthermore, misinformation about notions of fairness can have significant legal implications.[1]

Computer scientists have largely focused on developing mathematical notions of fairness and incorporating them in fielded ML systems. A much smaller collection of studies has measured public perception of bias and (un)fairness in algorithmic decision-making. However, one major question underlying the study of ML fairness remains unanswered in the literature: *Does the general public understand mathematical definitions of ML fairness and their behavior in ML applications?* We take a first step towards answering this question by studying non-expert comprehension and perceptions of one popular definition of ML fairness, *demographic parity* [5]. Specifically, we developed an online survey to address the following: **(1)** Does a non-technical audience comprehend the definition and implications of demographic parity? **(2)** Do demographics play a role in comprehension? **(3)** How are comprehension and sentiment related? **(4)** Does the application scenario affect comprehension?

We present participants ($n = 147$) with one of three simple, but realistic, decision-making scenarios where fairness plays a role – **Art Project (AP):** distributing awards for art projects amongst primary school students, **Employee Awards (EA):** distributing employee awards at a sales company, and **Hiring (HR):** distributing job offers to applicants. Each scenario is accompanied by a *fairness rule* (corresponding to demographic parity), expressed in each scenario's context. We ask several questions related to the participants' comprehension of and sentiment towards this rule. Tallying the number of correct responses to the comprehension questions gives us a *comprehension score* for each participant.

We find that this comprehension score is a consistent and reliable indicator of understanding demographic parity. Exploratory analysis reveals that education level is an important predictor for comprehension, and that *negative* sentiment is associated with *greater* comprehension of demographic parity. Moreover, the nature of the scenario (AP, EA, or HR) does not appear to influence comprehension. These findings inspire several areas for future work. Moreover, our work could be extended to similar investigation of other fairness definitions such as equal opportunity [6], equalized odds [6], calibration [8], and causal fairness [7].

[1]https://www.cato.org/blog/misleading-veritas-accusation-google-bias-could-result-bad-law

AIES '20, February 7–8, 2020, New York, NY, USA

ACM ISBN 978-1-4503-7110-0/20/02.
https://doi.org/10.1145/3375627.3375819

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; • **Applied computing** → *Law, social and behavioral sciences.*

KEYWORDS

human-computer interaction, algorithmic bias, fair machine learning, empirical study

ACKNOWLEDGMENTS

This research was supported by NSF IIS #1844462, NSF IIS #1844518, NSF IIS RI CAREER Award #1846237, and a gift from Google.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May 23* (2016).

[2] Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research)*, Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, New York, NY, USA, 149–159.

[3] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[4] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies* 2015, 1 (2015), 92–112.

[5] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*. ACM, New York, NY, USA, 214–226.

[6] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Info. Processing Sys. (NeurIPS'16)*. Curran Assoc. Inc., USA, 3323–3331.

[7] Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., USA, 4066–4076.

[8] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. On fairness and calibration. In *Proceedings of the 31st International Conference on Neural Info. Processing Sys. (NeurIPS'17)*. Curran Assoc. Inc., USA, 5684–5693.