Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics

BO COWGILL and FABRIZIO DELL'ACQUA, Columbia Business School
SAMUEL DENG, DANIEL HSU, NAKUL VERMA, and AUGUSTIN CHAINTREAU, Columbia Department of Computer Science

Why do biased algorithmic predictions arise, and what interventions can prevent them? We examine this topic with a field experiment about using machine learning to predict human capital. We randomly assign ≈400 AI engineers to develop software to predict standardized test scores of OECD residents under different experimental conditions. We then assess the resulting predictive algorithms using the realized test performances, and through randomized audit-like manipulations of algorithmic inputs. We also used the diversity of our subject population to measure whether demographically non-traditional engineers were more likely to notice and reduce algorithmic bias, and whether algorithmic prediction errors are correlated within programmer demographic groups. This document describes our experimental design and motivation; the full results of our experiment are available at https://ssrn.com/abstract=3615404.

CCS Concepts: • **Social and professional topics → Codes of ethics**; • **General and reference → Empirical studies**; **Experimentation**.

Additional Key Words and Phrases: Algorithmic fairness, field experiment.

ACM Reference Format:
Bo Cowgill, Fabrizio Dell'Acqua, Samuel Deng, Daniel Hsu, Nakul Verma, and Augustin Chaintreau. 2020. Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics. In *Proceedings of the 21st ACM Conference on Economics and Computation (EC '20), July 13–17, 2020, Virtual Event, Hungary.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3391403.3399545

1 EXTENDED ABSTRACT

Research about algorithmic fairness has grown enormously in recent years, along with efforts to raise awareness, technical and ethical knowledge about algorithmic systems. However, few studies have attempted to evaluate, audit or learn from these interventions or to connect them back to theory. This paper aims to step in that direction.

In this paper, we report the results of a large field experiment in AI development. Our subjects are ≈400 AI engineers. They are not told they are part of a field experiment and are recruited under a separate cover story to develop an algorithm for screening job candidates [5, 10].

Unbeknownst to these engineers, the details of their assigned tasks have been randomly altered as part of a field experiment. The experiment is constructed to test a series of hypotheses about why algorithmic bias arises and what managerial or policy interventions lead to its reduction. Our experiment gives us a direct view of algorithmic prediction technology while it is being assembled by AI engineers. Our setting creates measurement opportunities that would be impossible for learning processes in other economic settings.

EC '20, July 13–17, 2020, Virtual Event, Hungary
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-7975-5/20/07.
https://doi.org/10.1145/3391403.3399545

Our paper is related to economic approaches to studying algorithmic fairness [8, 9]. However, it is distinct in its focus on policy interventions in algorithmic production. A similar experiment [6] examines how managerial decision-making reacts to algorithmic fairness interventions.

1.1 Task

All subjects in our experiment were assigned the same job: Develop an algorithm to predict math performance from biographical features on a job application, and apply it to 20,000 new individuals who do not appear in the training data. For reasons we discuss in our full paper, math is an attractive topic for empirical studies of algorithmic bias.

As training data, engineers were given a sample of the OECD's *Programme for the International Assessment of Adult Competencies* ("PIAAC," [11]) dataset. PIAAC is the canonical dataset for cross-country and within country comparisons of numeracy and skills. The PIAAC data solves several critical research challenges for algorithmic bias researchers. To our knowledge, this is the first paper to utilize the PIAAC dataset in computer science research about algorithms. To facilitate others utilizing this dataset, we have made a cleaned and merged copy with documentation for other academic researchers.[1]

1.2 Hypotheses and Experimental Treatment Arms

Our experimental design included two benchmarks. The first was a control group in which engineers were given PIAAC data featuring realistic sample selection problems in the training data. As a second benchmark, we also gave a second randomly-selected group PIAAC data featuring no sample selection problems. In our remaining two treatments, engineers were given the first group's training data (featuring sample selection problems). However, these experimental groups were also given policy interventions. The first was given a non-technical reminder about the possibility of algorithmic bias. The other was given this reminder as well as a simple, jargon-free white paper about sample selection correction methods in machine learning [3, 4, 12].

1.2.1 Subtreatments: Performance Incentives. Within each of the above treatments, subjects were randomized into varying performance incentive schemes. The goal of this randomization was to measure the effectiveness of using incentives to reduce algorithmic bias.

1.2.2 Audit Manipulations. In addition, we randomized some aspects of the test data. Some engineers were given test data in which the same job candidate observation appeared twice, with the subject's gender (or other biographical details) flipped. This manipulation reveals how much the engineers' experimental algorithms explicitly use gender (or other biographical variables) in algorithmic forecasts.

This manipulation of inputs is a digital equivalent of resume-audit style research designs [2], in which researchers send fictitious resumes to employers with randomized content. Because our design features randomization both on screeners (programmers) and on candidates (subjects evaluated by the resume), our design resembles the "two-sided audit" design [1, 7].

1.2.3 Engineer Demographics. Our subject population contained substantial variation in gender, race and other demographic characteristics. We utilized this diversity to measure whether demographically non-traditional programmers were more likely to notice and reduce algorithmic bias, and whether prediction errors were correlated within demographic groups.

Complete Findings: For results, see our full manuscript at https://ssrn.com/abstract=3615404.

[1] https://tinyurl.com/piaac.

REFERENCES

[1] Amanda Agan, Bo Cowgill, and Laura Gee. 2019. The Effects of Salary History Bans: Evidence from a Field Experiment. *Working paper* (2019).

[2] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review* 94, 4 (2004), 991–1013.

[3] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[4] Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. 2008. Sample selection bias correction theory. In *International conference on algorithmic learning theory*. Springer, 38–53.

[5] Bo Cowgill. 2019. Automating Judgement and Decisionmaking: Theory and Evidence from Résumé Screening. *Working paper* (2019).

[6] Bo Cowgill, Fabrizio Dell'Acqua, and Sandra Matz. 2020. The Managerial Effects of Algorithmic Fairness Activism. In *AEA Papers and Proceedings*, Vol. 110. 85–90.

[7] Bo Cowgill and Patryk Perkowski. 2019. Agency and Homophily: Evidence from a Two-Sided Audit. *Working Paper* (2019).

[8] Bo Cowgill and Catherine E Tucker. 2019. Economics, fairness and algorithmic bias. *preparation for: Journal of Economic Perspectives* (2019).

[9] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. *Human decisions and machine predictions.* Technical Report. National Bureau of Economic Research.

[10] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 469–481.

[11] Andreas Schleicher. 2008. PIAAC: A new strategy for assessing adult competencies. *International Review of Education* 54, 5-6 (2008), 627–650.

[12] Bianca Zadrozny, John Langford, and Naoki Abe. 2003. Cost-Sensitive Learning by Cost-Proportionate Example Weighting.. In *ICDM*, Vol. 3. 435.