Ethical Algorithm Design

MICHAEL KEARNS
University of Pennsylvania
and
AARON ROTH
University of Pennsylvania

In this letter, we summarize the research agenda that we survey in our recent book *The Ethical Algorithm*, which is intended for a general, nontechnical audience. At a high level, this research agenda proposes *formalizing* the ethical and social values that we want our algorithms to maintain — values including privacy, fairness, and explainability — and then to embed these social values directly into our algorithms as part of their design. This broad research area is most mature in the area of privacy, specifically *differential privacy*. It is off to a good start in emerging areas like algorithmic fairness, and seems promising for more nebulous goals like explainability, if only we can find the right definitions. Most work in this area to date analyzes algorithms as isolated components, but game-theoretic and economic analysis will become increasingly important as we try and study the effects of algorithmic interventions in larger sociotechnical systems.

Categories and Subject Descriptors: F.0 [**Theory of Computation**]: General

General Terms: Fairness, Privacy, Explainability, Game Theory

Additional Key Words and Phrases: Fairness, Privacy, Game Theory

In the past decade, as machine learning has become more capable, it has also been deployed in increasingly consequential domains. Trained machine learning models are now used to automatically make credit and lending decisions [Koren 2016], to inform hiring and compensation decisions [Miller 2015], to inform bail and parole decisions [Angwin et al. 2016], and to help target healthcare resources [Obermeyer et al. 2019]. It is therefore not surprising that there has been rising concern over the potential for these technologies to violate basic social norms like fairness, privacy, transparency, and accountability [O'Neil 2016; Eubanks 2018; Schneier 2015; Benjamin 2019]. After all, when we have humans in important decision-making pipelines, we expect them to respect these basic social values. There is no reason we shouldn't also expect this of algorithmic decision-making.

How do we go about making sure that algorithms (often predictive models trained via machine learning) are *privacy preserving* or *fair*? An important thesis of our book [Kearns and Roth 2019] is that although it is necessary for the software engineers and scientists who are designing and deploying algorithms to themselves be ethical, it is far from sufficient. The vast majority of documented instances of algorithmic bias, for example, do not seem to be the result of any ill intentions on the part of the designers, but were instead the *unanticipated consequences* of applying the standard tools and principles of machine learning. This standard framework can run afoul in a number of ways. For example, because machine learning is data-

Authors' addresses: mkearns@cis.upenn.edu, aaroth@cis.upenn.edu

intensive, it can be tempting to use datasets of convenience — those with large numbers of data points consisting of cheaply and easily collected measurements. Such datasets often involve proxies for what we actually want to predict, instead of the real underlying values: arrests as a proxy for crime in criminal risk prediction applications [Berk et al. 2018], or medical costs as a proxy for severity of disease in healthcare applications [Obermeyer et al. 2019]. These proxies can reflect existing human and structural biases, which will in turn be mimicked in the learned model. For example, Black citizens being policed at higher rates may lead to artificially elevated arrest rates as a function of criminal activity, and having less access to healthcare can result in artificially low health costs as a function of disease severity. There is no reason to believe such biases in the data will be "unlearned" when applying machine learning.

Another basic problem is that high-dimensional model optimization often results in "corner solutions" that are difficult to understand, and cannot be expected to satisfy any property that you didn't explicitly ask for (like some notion of fairness or privacy). We can be confident of little about a trained model other than that it will perform well according to the objective function that was optimized for, which is usually some narrow and myopic proxy for aggregate error or profit. This manifests itself frequently. For example, standard machine learning methods often lead models to "memorize" training data points in ways that let others recover them (an unanticipated privacy violation) [Shokri et al. 2017], and to have higher error on minority populations (an unanticipated fairness violation [Angwin et al. 2016]). In fact, optimizing average error naturally leads to models that disadvantage minorities, simply because (by definition) minorities contribute less to the overall error. Thus whenever an algorithm cannot simultaneously fit two different subsets of the population optimally, it will fit the majority class at the expense of the minority class. And seemingly sensible precautions — such as hiding sensitive demographic information from the algorithm — may exacerbate this problem by forcing it to attempt to fit different populations with the same model.

If the problem is that machine learning produces models that are unpredictable once we move away from their defined objectives and constraints, the solution is to be explicit about what exactly we want from our algorithms when we ask that they be private or fair. This is a difficult task, because words like "privacy" and "fairness" can have a broad set of informal and nuanced meanings, and can mean very different things to different people. Language and law are often able to skirt such ambiguity by deferring many precise questions to human beings and courts when necessary. But constraining the sorts of optimization procedures used in machine learning requires mathematical rigor and definitions. Coming up with definitions that capture at least some parts of the intuitive concepts is difficult.

But formalization is not impossible. A success story in this broad research agenda is differential privacy [Dwork et al. 2016], which is the focus of the first chapter of our book. Differential privacy does not encompass everything that is meant by the word "privacy", as no single formal definition could. But it does capture much of what one might mean by privacy in statistical computations, and provides a language in which to discuss different kinds of privacy guarantees. Informally, differential privacy is a guarantee of plausible deniability: there should be no statistical procedure that can determine (substantially) better than random guessing whether a particular individual's data was used in a computation or not. Thus (almost) nothing can be learned about them from the output of a computation that could not have been learned absent their data. This provides a way to disentangle the "secrets" of a particular person (information that can only be learned by examining a person's data, which differential privacy protects) from information that is implicit in population-level correlations (which differential privacy does not protect). Because it is a formal guarantee, we can design algorithms that *provably* provide differential privacy — and we can study the inevitable tradeoffs that arise between privacy and accuracy. Moreover, differential privacy is parameterized quantitatively: the words "substantially" and "almost" in our informal discussion above are precisely quantified in its formal definition. This allows us to think quantitatively about Pareto frontiers: there are different ways to trade off accuracy with privacy that we can map out using the language of differential privacy. *How* we want to make the tradeoff is a policy decision that has no universal answer — but differential privacy provides a language in which to have the debate. And in the decade and a half that has passed since its original introduction, differential privacy has gone from an object of mathematical study to a real technology that is becoming widely deployed in both industry and government.

The fairness in machine learning literature (the focus of our second chapter) is at a nascent stage in which there is no broad agreement on what the right definitions are, and our understanding of the properties of the definitions we have remains elementary [Chouldechova and Roth 2020]. Nevertheless, we can aspire to achieve for algorithmic fairness what differential privacy has achieved: to develop a formal language in which to precisely discuss different kinds of fairness guarantees, to study what is (and is not) achievable subject to these constraints, and to eventually translate these guarantees from the whiteboard to deployed technology. This won't be easy — it will be a decades-long research agenda, as it was (and remains) for data privacy. And we already know that the study of fairness will be messier than the study of privacy, in the sense that there are multiple statistical fairness constraints that one could ask for which are mutually incompatible with one another [Kleinberg et al. 2016; Chouldechova 2017]. There is also a divide between "statistical" notions of fairness that are already practical but provide limited promises to individuals, and definitions that provide stronger individual fairness guarantees [Dwork et al. 2012; Joseph et al. 2016] but which have various barriers to realistic implementation. One promising line of work aims to find definitions that can bridge this gap in various ways, being achievable without needing to make unrealistic assumptions on the data, but still corresponding to a promise to individuals [Kearns et al. 2018; Hebert-Johnson et al. 2018; Gillen et al. 2018; Dwork et al. 2019; Sharifi-Malvajerdi et al. 2019; Kim et al. 2018; Yona and Rothblum 2018; Jung et al. 2019; Ilvento 2019].

Another shortcoming of the aforementioned science is that it tends to focus myopically on isolated machine learning problems, whereas fairness issues generally do not arise in a vacuum around a single algorithm, but rather as part of larger sociotechnical systems in which algorithms are deployed. Data is generated by *people*, and the decisions of algorithms have effects in the world that change human incentives and behavior. Understanding how algorithms affect larger-scale societal dynamics and equilibrium is important to understanding fairness more broadly, and is the place where algorithmic game theory (the topic of our third chapter) has broad potential to be useful. There is already some early work studying the effects of various fairness constraints in dynamic settings (both game theoretic and not) [Jabbari et al. 2017; Hu and Chen 2018; Liu et al. 2018; Kannan et al. 2019; Liu et al. 2020; Jung et al. 2020], but much remains to be done here.

While the study of algorithmic fairness may be a decade or so behind data privacy, it is off to a promising start in the sense that it is at least already grappling with *definitions*. Another major theme of our book is that precise definitions are a crucial prerequisite to progress, especially when trying to pin down nebulous social values in ways that can be embedded in code. How can we make progress on accomplishing our goals if we cannot yet enunciate what our goals are? A lesson from both algorithmic privacy and algorithmic fairness is that the process of working through definitions carefully can make clear why seemingly sensible heuristics often fail. Attempting to "anonymize" data is doomed to failure because of the prevalance of external sources of information; instead we need to directly constrain what can be inferred from outputs (as differential privacy does). Hiding demographic information from machine learning algorithms fails — both because it doesn't work (it is distressingly easy to predict most demographic attributes from seemingly innocuous traits) — and because it can actually exacerbate error disparities in algorithms by forcing them to fit different distributions with the same model. Instead we must directly constrain the outcomes of the algorithm (which most current fairness definitions do). In our fifth chapter we briefly discuss important parts of this research agenda — like "explainability" and "interpretability" — that so far seem still to be awaiting the right definitions. We might be primed to make rapid progress on these important goals if only we can figure out what they should mean.

Our book is designed to be approachable to a general, non-technical reader, but we hope that it will also be informative and enjoyable for researchers interested in the intersections of machine learning, algorithms, game theory and related topics. While it is written without equations, we attempt to communicate key ideas in each research area without undue dilution. We expect that most readers will find at least some topics that are new to them: differential privacy (Chapter 1), fairness in machine learning (Chapter 2), algorithmic game theory (Chapter 3), adaptive data analysis (Chapter 4), and important topics that have yet to be formalized in convincing ways — such as transparency, explainability, accountability, and AI safety (Chapter 5).

REFERENCES

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. 2016. Machine bias. *Propublica*.

Benjamin, R. 2019. *Race after technology: Abolitionist tools for the new jim code.* John Wiley & Sons.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. 2018. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 0049124118782533.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data 5,* 2, 153–163.

Chouldechova, A. and Roth, A. 2020. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM 63,* 5, 82–89.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

Dwork, C., Kim, M. P., Reingold, O., Rothblum, G. N., and Yona, G. 2019. Learning from outcomes: Evidence-based rankings. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 106–125.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. 2016. Calibrating noise to sensitivity in private data analysis. *Journal of Privacy and Confidentiality 7,* 3, 17–51.

Eubanks, V. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press.

Gillen, S., Jung, C., Kearns, M., and Roth, A. 2018. Online learning with an unknown fairness metric. In *Advances in Neural Information Processing Systems*. 2600–2609.

Hebert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. 2018. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*. 1939–1948.

Hu, L. and Chen, Y. 2018. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*. 1389–1398.

Ilvento, C. 2019. Metric learning for individual fairness. *arXiv preprint arXiv:1906.00250*.

Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., and Roth, A. 2017. Fairness in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 1617–1626.

Joseph, M., Kearns, M., Morgenstern, J. H., and Roth, A. 2016. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*. 325–333.

Jung, C., Kannan, S., Lee, C., Pai, M. M., Roth, A., and Vohra, R. 2020. Fair prediction with endogenous behavior. In *The Twenty-First ACM Conference on Economics and Computation*.

Jung, C., Kearns, M., Neel, S., Roth, A., Stapleton, L., and Wu, Z. S. 2019. Eliciting and enforcing subjective individual fairness. *arXiv preprint arXiv:1905.10660*.

Kannan, S., Roth, A., and Ziani, J. 2019. Downstream effects of affirmative action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 240–248.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*. 2564–2572.

Kearns, M. and Roth, A. 2019. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press.

Kim, M. P., Reingold, O., and Rothblum, G. N. 2018. Fairness through computationally-bounded awareness. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. 4847–4857.

Kleinberg, J., Mullainathan, S., and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Koren, J. R. 2016. What does that web search say about your credit? *Los Angeles Times*. Retrieved 9/15/2016.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. 2018. Delayed impact of fair machine learning. *arXiv preprint arXiv:1803.04383*.

Liu, L. T., Wilson, A., Haghtalab, N., Kalai, A. T., Borgs, C., and Chayes, J. 2020. The disparate equilibria of algorithmic decision making when individuals invest rationally. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 381–391.

Miller, C. C. 2015. Can an algorithm hire better than a human? *The New York Times*. Retrieved 4/28/2016.

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

O'Neil, C. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Schneier, B. 2015. *Data and Goliath: The hidden battles to collect your data and control your world*. WW Norton & Company.

Sharifi-Malvajerdi, S., Kearns, M. J., and Roth, A. 2019. Average individual fairness: Algorithms, generalization and experiments. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds. 8240–8249.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 3–18.

Yona, G. and Rothblum, G. 2018. Probably approximately metric-fair learning. In *International Conference on Machine Learning*. 5680–5688.