Translation Tutorial: Fairness and Friends

Falaah Arif Khan
New York University
fak.723@gmail.com

Eleni Manis
Surveillance Technology Oversight Project
em579@nyu.edu

Julia Stoyanovich
New York University
stoyanovich@nyu.edu

ABSTRACT

Recent interest in codifying *fairness* in Automated Decision Systems (ADS) has resulted in a wide range of formulations of what it means for an algorithm to be "fair." Most of these propositions are inspired by, but inadequately grounded in, scholarship from political philosophy. This tutorial aims to correct that deficit. We critically evaluate different definitions of fairness by contrasting their conception in political philosophy (such as Rawls's fair equality of opportunity or formal equality of opportunity) with the proposed codification in Fair-ML (such as statistical parity, equality of odds, accuracy) to provide a clearer lens with which to view existing results and to identify future research directions. A key novelty of this tutorial is the use of technical artwork to make ideas more relatable and accessible, based on our ongoing work on a responsible data science comic book series, available at https://dataresponsibly.github.io/comics/.

ACM Reference Format:
Falaah Arif Khan, Eleni Manis, and Julia Stoyanovich. 2021. Translation Tutorial: Fairness and Friends. In *Proceedings of FAT*ML '21: ACM Conference on Fairness, Accountability and Transparency in Machine Learning (FAccT '21).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXX/XXXXXXX

1 INTRODUCTION

Automated Decision Systems (ADS) are broadly used socio-technopolitical systems [19], and codifying 'fairness' in the context of these systems requires a harmonization of scholarship in machine learning, political philosophy and law. The primary goal of our tutorial is to ground current approaches in fair machine learning (fair-ML) in a better understanding of their counterparts in political philosophy. While most propositions of fair-ML draw on scholarship from political and moral philosophy, a critical survey of literature indicates both a naïve understanding of philosophical theories and a misunderstanding of their applicability to real-world contexts. Another dimension of complexity comes from the legal doctrines that influence the conception of 'fairness' definitions that will hold up against the rule of law. The goal of our tutorial is to distill the influence from these three fields, helping ground current and future fair-ML scholarship [6].

A deeper reading of the justice literature in political philosophy illustrates the limitations of current formulations of fairness in ML. For example, a popular formulation casts Fairness as Equality of Opportunity (EoP) using economists' models of equality of opportunity [12]. However, by posing the fairness task as a mapping from circumstance and effort directly to utility, this formulation bypasses equality of opportunity to achieve equality of outcomes. We trade economic conceptions of EOP for philosophical conceptions, setting the stage for a more plausible mapping of statistical fairness measures onto EOP conceptions. We then consider how political philosophy can provide normative guidance as to which statistical notion of fairness is applicable in which context.

Our presentation introduces conceptions of EOP in a philosophical context. This allows us to identify and organize justice considerations that the fair-ML literature currently overlooks or underemphasizes. For example, Rawls supplements his EOP principle with a principle guaranteeing equal rights and liberties including freedom of speech and freedom of association. Consider this in the context of "fair" hiring of people with disabilities: "disability" would be treated as a protected class and removed from consideration, but algorithms could still infer disability from other proxy variables. If social media information is used to infer disability status—for example, based on membership in certain social groups or on posting about disability-related issues—then a scheme that discriminates on the basis of "inferred" disability would incentivize people against joining such groups and speaking about such topics. Such algorithms could be fair, but fundamentally unjust: they would violate a commitment to equal basic rights and liberties such as freedom of speech, and freedom of association.

We also present a reinterpretation of recent impossibility results, in light of the highlighted limitations. For example, much of the recent fair-ML literature has been working to formulate Fairness as Justice (Libertarian/Rawlsian/Roemerian), but what has actually been codified is only the EoP principle of Justice. We interpret recent impossibility results [5, 10, 13] to demonstrate the mutual incompatibility of Formal and Substantive EoP and argue that fully satisfying either ideal would be problematic in itself.

We will conclude by using these insights to propose future directions for research in Fair-ML, while calling out the limitations in the guidance from political philosophy.

A key novelty of this tutorial is the use of technical artwork to depict ideas and make the content more relatable to the audience. For example, the comic panel describing the three-headed dragon of bias is shown in Figure 1. All artwork used in this tutorial is taken from the upcoming Vol. 2 of the Data, Responsibly comic books, currently under development by the authors, see https://dataresponsibly.github.io/comics/.

2 IMPACT

There is a critical lack of guidance from political philosophy in fair-ML scholarship, and our tutorial is a robust attempt at bridging this gap. We hope to provide a grounding for future research in fair-ML,

FAccT '21, June 03–05, 2021, online

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXX/XXXXXXX

Figure 1: The three-headed dragon of pre-existing, technical, and emergent bias.

and in responsible data science more broadly, and to invite scholars from complementary fields such as law and ethics to continue this critical dialogue.

Further, with this tutorial we hope to pave the way for new modes of scholarship beyond the research papers. Mediums such as technical artwork, graphic guides, and comic books [2–4] are a rich but underrepresented source of scholarship. We hope to demonstrate their utility through our tutorial, and to popularize their adoption more broadly in the ML community.

3 TUTORIAL STRUCTURE

Our 90-minute tutorial will compare notions of fairness and justice in political philosophy to their counterparts in fair-ML scholarship and present lessons and future directions for research.

Bias in Algorithmic Systems [10min]. We will motivate the problem of enforcing 'fairness' in the outcomes of Automated Decision Systems (ADS) by defining what an ADS is (and is not), and explaining we mean by bias in the context of ADS [7, 11, 15, 19]. The idea that ADS are socio-techno-political systems will be a common thread across the tutorial, motivating scholars to think deeply about how to ground fair-ML scholarship in political philosophy and in the social sciences.

Fairness and Equality of Opportunity [25min]. Equality of opportunity [EOP] is a helpful frame for understanding fair-ML. However, fair-ML literature has not yet established philosophically grounded and intuitively compelling connections between different EOP views and different fairness concerns. Toward this end, we introduce versions of EOP and the philosophical views that ground them: Libertarianism, Formal Equality of Opportunity, and Substantive Equality of Opportunity (Rawlsian and Luck Egalitarian) [1, 17, 18, 21]. We distinguish between equality of developmental opportunities, EOP over a lifetime, and EOP at a decision point (fair-ML's focus) to make better sense of substantive EOP [9]. We also introduce philosophically well-known objections to achieving EOP. This section focuses on clarifying and providing context for philosophical ideas that are already influential in ML scholarship.

Fairness Definitions in Machine Learning [15min]. We will present a critical review of the definitions of 'fairness' that have been proposed in ML [5, 8, 12, 20]. Throughout, concepts will be presented using light and relatable artwork to make the tutorial more accessible to people of different backgrounds and technical abilities.

Fair But Not Just and Other Limitations [20min]. The exposition of fairness notions from political philosophy and fair-ML will be followed by a critical discussion on gaps between philosophical theories, their technical codification, and legal doctrines. With the grounding provided in the preceding section, we will also highlight the source of this lapse, including a naïve application of the underlying theory, an imprecise technical definition, lack of guidance in political philosophy, or incompatibility with legal doctrines. We will use practical examples to demonstrate the harm of reductive formulations including the infamous COMPAS tool, and its (less well known) 'COMPAS Women' and 'COMPAS Youth' counterparts. Another practical example is the impact of hiring ADS on people with disabilities, where we will demonstrate the perils of current formulations that would give rise to decisions that are at the same time 'fair' (i.e., they satisfy some parity conditions) and but 'unjust' (i.e., they infringe on the basic rights and liberties of applicants).

Reinterpreting Impossibility Results [10min]. There is much fair-ML scholarship on the mutual incompatibility of fairness measures [5, 10, 13, 14, 16**?**]. We will interpret this incompatibility in terms of the conflict between formal EOP and substantive EOP, which also pull mutually incompatible directions. We propose a reinterpretation of the recent impossibility results as evidence of the limitations discussed in the preceding section.

Lessons and Proposed Directions for Research [10min]. We conclude with an overview of 'fairness' in different philosophical views and legal doctrines, and highlight open areas for future work. We underscore that entire areas and formulations such as developmental EoP seem to be overlooked in favor of discrete point interpretations (i.e., results of a competition view) in fair-ML. Further, we will stress that reductive formulations can arise due to a misinterpretation or naïve reading of political philosophy doctrines that emboldens their application in spheres in which theory provides little to no guidance.

4 PRESENTER BIOS

Falaah Arif Khan is an Artist-in-Residence at the Center for Responsible AI at New York University. She has been a Teaching Assistant at Shiv Nadar University & a speaker at the TechAide AI4Good Conference+Hackathon '20, the Algorithmic Justice Webinar Series by Rutgers Law School & at Dell's Sparks Tech Forum. Links to some talks are at https://falaaharifkhan.github.io/research/.

Eleni Manis is the Research Director at the Surveillance Technology Oversight Project. Eleni began her career as an Assistant Professor of Philosophy at Franklin & Marshall College, where she taught courses ranging from History of Political Philosophy and Ethics to Contemporary Political Philosophy and seminars on Law & Philosophy, Democracy, and Global Justice.

Julia Stoyanovich is an Assistant Professor of Computer Science and of Data Science at New York University, where she directs the Center for Responsible AI. She has been teaching courses in the academic setting since 2009 and regularly speaks to a variety of audiences about responsible data science. A list of her recent courses and talks is at https://dataresponsibly.github.io.

REFERENCES

[1] Elizabeth S. Anderson. 1999. What Is the Point of Equality? *Ethics* 109, 2 (1999), 287–337. http://www.jstor.org/stable/10.1086/233897

[2] Falaah Arif Khan and Abhishek Gupta. 2020. *Decoded Reality.* https://ai-ethics.github.io/decoded-reality/intro.html.

[3] Falaah Arif Khan and Zachary C Lipton. 2020. *Vol 1: Machine Learning Yearning.* Vol. 1. Superheroes of Deep Learning. https://github.com/acmi-lab/superheroes-deep-learning.

[4] Falaah Arif Khan and Julia Stoyanovich. 2020. *Mirror, Mirror.* Vol. 1. Data Responsibly Comics. https://dataresponsibly.github.io/comics/.

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

[6] Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* 63, 5 (2020), 82–89. https://doi.org/10.1145/3376898

[7] Kimberle Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum* 1 (1989), 139–167.

[8] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, Shafi Goldwasser (Ed.). ACM, 214–226. https://doi.org/10.1145/2090236.2090255

[9] Joseph Fishkin. 2014. *Bottlenecks: A New Theory of Equal Opportunity.* Oup Usa.

[10] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. *CoRR* abs/1609.07236 (2016). arXiv:1609.07236 http://arxiv.org/abs/1609.07236

[11] Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. *ACM Trans. Inf. Syst.* 14, 3 (1996), 330–347. https://doi.org/10.1145/230538.230561

[12] Hoda Heidari, Michele Loi, Krishna P. Gummadi, and Andreas Krause. 2019. A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019.* ACM, 181–190. https://doi.org/10.1145/3287560.3287584

[13] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. *CoRR* abs/1609.05807 (2016). arXiv:1609.05807 http://arxiv.org/abs/1609.05807

[14] Sendhil Mullainathan and Marianne Bertrand. 2004. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review* 94 (02 2004), 991–1013. https://doi.org/10.2139/ssrn.422902

[15] S.U. Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism.* NYU Press. https://books.google.co.in/books?id=g8OSDgAAQBAJ

[16] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366 (10 2019), 447–453. https://doi.org/10.1126/science.aax2342

[17] JOHN RAWLS. 1971. *A Theory of Justice.* Harvard University Press. http://www.jstor.org/stable/j.ctvjf9z6v

[18] John Roemer. 2002. Equality of opportunity: A progress report. *Social Choice and Welfare* 19, 2 (2002), 455–471. https://EconPapers.repec.org/RePEc:spr:sochwe:v:19:y:2002:i:2:p:455-471

[19] Julia Stoyanovich, Bill Howe, and H. V. Jagadish. 2020. Responsible Data Management. *Proc. VLDB Endow.* 13, 12 (2020), 3474–3488. http://www.vldb.org/pvldb/vol13/p3474-stoyanovich.pdf

[20] Ke Yang, Joshua Loftus, and Julia Stoyanovich. 2020. Causal intersectionality for fair ranking. (06 2020).

[21] H.P. Young and Russell Sage Foundation. 1994. *Equity: In Theory and Practice.* Princeton University Press. https://books.google.co.in/books?id=XVK5AAAAIAAJ