Economic Theories of Distributive Justice for Fair Machine Learning

Krishna P. Gummadi
MPI-SWS, Saarbrücken
gummadi@mpi-sws.org

Hoda Heidari
ETH Zürich
hheidari@inf.ethz.ch

ABSTRACT

Machine Learning is increasingly employed to make consequential decisions for humans. In response to the ethical issues that may ensue, an active area of research in ML has been dedicated to the study of algorithmic unfairness. This tutorial introduces fair-ML to the web conference community and offers a new perspective on it through the lens of the long-established economic theories of distributive justice. Based on our past and ongoing research, we argue that economic theories of equality of opportunity, inequality measurement, and social choice have a lot to offer—in terms of tools and insights—to data scientists and practitioners interested in understanding the ethical implications of their work. We overview these theories and discuss their connections to fair-ML.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning**; • **Applied computing** → **Economics**.

KEYWORDS

social choice theory; equality of opportunity; inequality

ACM Reference Format:
Krishna P. Gummadi and Hoda Heidari. 2019. Economic Theories of Distributive Justice for Fair Machine Learning. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3320101

1 INTRODUCTION

Automated data-driven decision-making tools are increasingly employed to make consequential decisions for human subjects— examples include employment [19], credit lending [20], criminal justice [1], policing [23], and medicine [14]. Decisions made in this fashion can have a long-lasting impact on society and may affect certain individuals or groups negatively [25]. This realization has recently spawned a new active area of research into quantifying and guaranteeing fairness for machine learning [4, 12, 17].

As concerns over algorithmic unfairness and discrimination continue to grow in magnitude and depth, it is timely—and critical—for data scientists and practitioners to be armed with a toolbox of models and mechanisms to quantify and tackle algorithmic unfairness in their application domains. This tutorial offers a new perspective on fair ML through the lens of the long-established *economic* theories of distributive justice. The extensive economic literature on *equality of opportunity*, *measurement of inequality*, *social choice theory*, and *fair division* has a lot to offer—in terms of models and tools—to data scientists and practitioners, interested in understanding the ethical implications of their work.

The main objectives of this tutorial are:

- to overview the growing line of work on fairness for ML;
- to survey the economic theories of distributive justice;
- to put Fair ML into economic perspective and terminology;
- to cast the well-established economics of distributive justice as a blueprint to guide and inform future research into fairness for Machine Learning.

We begin with an overview of the fair ML literature. We introduce existing notions of algorithmic (un)fairness and summarize some of the major themes and findings in Fair ML. We then cast these notions as special cases of economic models of Equality of Opportunity (EOP). Through this lens, we offer a better understanding of the moral assumptions underlying technical definitions of fairness. Second, we discuss the conception of unfairness as inequality. We overview the axiomatic characterization of measures of (income) inequality and present them as a unifying framework for quantifying both individual- and group-level unfairness. Third, we discuss the "leveling down objection" to equality. We propose the use of cardinal social welfare functions to address this issue and as an efficient method for bounding inequality. Last but not least, we discuss how differing (group) preferences can justify unequal outcomes, drawing on the concepts of envy and equity from fair allocation.

2 SCOPE

This tutorial discusses several seminal papers from economics (e.g., [2, 3, 13, 18]) and connects them to the growing body of work on fairness for ML. We overview of the recent research on algorithmic fairness. We introduce group-level notions of fairness—which require that given a classifier, a certain fairness metric is equal across all protected groups (see e.g. [17, 27, 28])—-as well as individual-level notions of fairness [9]—which requires that two individuals who are similar with respect to the task at hand receive similar classification outcomes.

Fair ML as Equality of Opportunity: Next, we offer a new moral framework for understanding these notions by mapping them to economic models of Equality of opportunity (EOP) [16]. EOP is a widely supported ideal of fairness, and it has been extensively studied in political philosophy and economics [22]. We show that through this conceptual mapping, many existing definitions of algorithmic fairness, such as predictive value parity and equality odds, can be interpreted as special cases of EOP. This approach

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3320101

allows us to explicitly spell out the moral assumptions underlying each notion of fairness. Moreover, it confers a moral meaning to the recent fairness impossibility results.

(Un)fairness as (In)equality: The understanding of fairness as some form of *equality* is the fundamental basis for many theories of justice. Given the distribution of individuals' *well-being* or *benefits*, measures of inequality capture the degree to which the distribution is unequal and allow for a direct comparison of inequality between various distributions. We start by noting that parity-based definitions of algorithmic fairness, such as statistical parity [5], disparate impact [10], and equality of opportunity [12], can be thought of as seeking to minimize some form of inequality—for different notions of *benefit* or *well-being*. This observation motivates our interest in inequality measurement. We overview the axiomatic characterization of inequality indices, using Gini, Theil, and Generalized entropy indices as examples [6, 8]. We discuss these axioms—such as *population-invariance* and *progressive transfer* principle—in the context of fair ML. We argue that inequality indices can be utilized to extend existing definitions of algorithmic fairness to multiple groups and settings beyond binary classification. Furthermore, a particular structural property of these indices, called *additive decomposability*, allows us to interpolate between individual and group-level (un)fairness and observe the tradeoffs between the two [24]. The main challenges we see in utilizing inequality indices as measures of algorithmic unfairness are (a) defining the right notion of well-being/benefit to equalize across groups/individuals, and (b) finding efficient and precise mechanisms for bounding inequality.

Next, we overview two economic theories of distributive justice that depart from the idea of fairness as equality: social choice theory and fair division. Social choice theory is concerned with aggregating individual preferences to pick a *just* collective outcome, where justice is defined through a set of axioms. Fair division is concerned with the division of a limited resource (e.g., a cake) among individuals with heterogeneous valuations/preferences.

Social Choice Theory: When individual utilities are interpersonally comparable, a cardinal social welfare function can be designed to choose a collective outcome [13]. Social welfare functions can be interpreted as measures of distributive justice behind a *veil of ignorance* [21]. This interpretation motivates our interest in employing them to measure algorithmic fairness. We go over the axiomatic characterization of cardinal social welfare functions, discussing such axioms as *symmetry* and *independence of unconcerned agents* in the context of fair ML. We next note that according to the Debreu-Gorman theorem [7, 11], the family of social welfare functions satisfying the above axioms is strikingly small. Furthermore and unlike measures of inequality, this class of social welfare functions enjoys a convex formulation. Therefore, it can be readily integrated into any convex loss minimization pipeline [15]. Last but not least, we briefly discuss the connections between welfare and inequality aversion through a welfare-based interpretation of the Atkinson's measure of inequality [3] and show that guaranteeing high social welfare usually leads to low inequality in practice.

Fair Division: Fair division accounts for the fact that different entities may have different preferences for different outcomes. We formally introduce the notions of *envy* and *equity*. An allocation is *envy-free* if no individual prefers the allocation of another to their own [18, 26]. We discuss one recent adaptation of no-envy to capture algorithmic fairness at the group level [29].

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *Propublica* (2016).
[2] Kenneth J. Arrow. 2012. *Social choice and individual values*. Vol. 12. Yale university press.
[3] Anthony B. Atkinson. 1970. On the measurement of inequality. *Journal of Economic Theory* 2, 3 (1970), 244–263.
[4] Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *California Law Review* (2016).
[5] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *ICDM workshops*. 13–18.
[6] Frank A. Cowell. 2000. Measurement of inequality. *Handbook of Income Distribution* 1 (2000), 87–166.
[7] Gerard Debreu. 1959. *Topological methods in cardinal utility theory*. Technical Report. Cowles Foundation for Research in Economics, Yale University.
[8] Bhaskar Dutta. 2002. Inequality, poverty and welfare. *Handbook of social choice and welfare* 1 (2002), 597–633.
[9] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the Innovations in Theoretical Computer Science Conference*. ACM, 214–226.
[10] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of KDD*. ACM, 259–268.
[11] William M. Gorman. 1968. The structure of utility functions. *The Review of Economic Studies* 35, 4 (1968), 367–390.
[12] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of NIPS*. 3315–3323.
[13] John C. Harsanyi. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of political economy* 63, 4 (1955), 309–321.
[14] Robert David Hart. 2017. If you're not a white male, artificial intelligence's use in healthcare could be dangerous. *Quartz* (July 2017).
[15] Hoda Heidari, Claudio Ferrari, Krishna P. Gummadi, and Andreas Krause. 2018. Fairness Behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making. In *Proceedings of NeurIPS*.
[16] Hoda Heidari, Michele Loi, Krishna P. Gummadi, and Andreas Krause. 2019. A Moral Framework for Understanding of Fair ML through Economic Models of Equality of Opportunity. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (ACM FAT*)*.
[17] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. In *In proceedings of the 8th Innovations in Theoretical Computer Science Conference*.
[18] Serge-Christophe Kolm. 2002. *Justice and equity*. MIT Press.
[19] Claire Miller. 2015. Can an Algorithm Hire Better than a Human? *The New York Times* (June 25 2015).
[20] Kevin Petrasic, Benjamin Saul, James Greig, and Matthew Bornfreund. 2017. Algorithms and bias: What lenders need to know. *White & Case* (2017).
[21] John Rawls. 2009. *A theory of justice*. Harvard university press.
[22] John E. Roemer and Alain Trannoy. 2015. Equality of opportunity. In *Handbook of income distribution*. Vol. 2. Elsevier, 217–300.
[23] Cynthia Rudin. 2013. Predictive Policing Using Machine Learning to Detect Patterns of Crime. *Wired Magazine* (August 2013).
[24] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual and Group Unfairness via Inequality Indices. In *Proceedings of KDD*.
[25] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Queue* 11, 3 (2013), 10.
[26] Hal R. Varian. 1974. Equity, envy, and efficiency. *Journal of economic theory* 9, 1 (1974), 63–91.
[27] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of WWW*. 1171–1180.
[28] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*.
[29] Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. 2017. From Parity to Preference-based Notions of Fairness in Classification. In *Proceedings of NIPS*. 228–238.