Fairness and Discrimination in Mechanism Design and Machine Learning

Jessie Finocchiaro[1] , **Roland Maio**[2] , **Faidra Monachou**[3] , **Gourab K Patro**[4] , **Manish Raghavan**[5] , **Ana-Andreea Stoica**[2] and **Stratis Tsirtsis**[6]

[1]CU Boulder, [2]Columbia University, [3]Stanford University, [4]IIT Kharagpur, [5]Cornell University, [6]Max Planck Institute for Software Systems

Abstract

As fairness and discrimination concerns permeate the design of both machine learning algorithms and mechanism design problems, we discuss differences in approaches between these two fields. We aim to bridge these two communities into a cohesive narrative that encompasses both the large-scale capabilities of machine learning and group-focused fairness as well as the strategic incentives and utility-based notions of fairness from mechanism design, showing their necessity in designing a fair pipeline.

1 Introduction

Fairness and equity are contested concepts. Paraphrasing [Dworkin, 2002], *"People who praise or disparage fairness disagree about what they are praising or disparaging."* Indeed, when computer scientists and economists are faced with the problem of discrimination, they still struggle with agreeing on the appropriate definition of fairness.

Both machine learning (ML) and mechanism design (MD) have developed frameworks for defining and applying fair systems in which a central planner optimizes a collective decision. While, historically, the fields have differed in the application domains and methodologies used, ML is increasingly applied in decision-making problems, in which MD is commonly used (e.g. advertisement targeting). More specifically, supervised learning systems are increasingly being used for *resource allocation*, with direct consequences on economic and other aspects of society. A prediction or classification is often tied to crucial decisions for which fairness becomes central, such as receiving parole and being hired, and mechanism design often uses the outputs of machine learning systems to make community-level decisions. Therefore, it is essential to bridge both the normative and methodological components that comprise fair and equitable systems in machine learning and mechanism design. This is the central purpose of this paper: to identify areas in which ML and MD have inspired each other in defining fair systems, and to fill in the 'gaps' that are lost in translating between ML and MD. While ML has started to incorporate notions of welfare in the design of classification algorithms, there is still a tension, in the form of trade-offs, between using parity-based notions and welfare notions from economics to define fairness. These trade-offs are often amplified in feedback loops that ML systems often involve. Conversely, while the field of mechanism design has focused on standard individual-level notions of equity, machine learning has expanded on these through group-level notions of disparity that can help diagnose large scale inequities, whether they occur organically in society or through the use of an algorithm. These developments arguably bring novel opportunities for the field of machine learning to learn and understand socioeconomic disparities in different communities.

Finally, we argue that mere translation between fairness metrics is also not enough: we need to understand the incentives that lead to their creation (as [Hu and Chen, 2020] and [Sen, 1979] ask, 'equality of what?') and to formulate a pipeline in which we can define contextual fairness notions that address the underlying dynamics of society, taking into account not only outcome distributions but also people's preferences and incentives. We conclude by discussing application domains in which both MD and ML fruitfully collaborate.

2 Differences between Mechanism Design and Machine Learning

Recent works in machine learning have opened a human-centric direction of the field, extending it from a *descriptive* nature (e.g. image classification) to a *prescriptive* one that automates human decision making. This shift has drawn attention to biases inherent to learning and predicting outcomes from historically prejudiced data [Angwin *et al.*, 2016; Buolamwini and Gebru, 2018; Barocas and Selbst, 2016]. Much of the fairness-related work in prescriptive ML relies on establishing parity conditions for legally protected groups, without considering welfare notions or strategic behavior. However, these notions lie at the core of mechanism design, which leverages individuals' strategies and sense of utility in establishing equilibrium solutions and Pareto-efficiency, but usually without conceptualizing the impact for different social groups. As [Abebe and Goldner, 2018;

Kasy and Abebe, 2020] point out, understanding the differences between the two fields and bridging different notions of fairness is essential in improving access to opportunity for different communities.

2.1 Fairness in machine learning

Multiple quantitative definitions of fair ML algorithms have been proposed; interestingly, their common characteristic seems to be that they agree to disagree. Some notions focus on *individual fairness*, where one wants to "treat similar inidividuals similarly," [Dwork *et al.*, 2012] while others prioritize *group fairness* in which the algorithm strives for the average treatment of members of different groups to be equal. Individual fairness imposes a much stronger constraint on what constitutes a fair treatment; an algorithm must be fair with respect to every single individual—not just on average. On the other hand, group fairness notions generally assess the bias of a system in a group through notions such as equalized odds, equal opportunity and demographic parity [Hardt *et al.*, 2016; Verma and Rubin, 2018; Mehrabi *et al.*, 2019], acting as a relaxation of individual fairness and assessing the large-scale effect of an algorithm on vulnerable populations.

2.2 Fairness in mechanism design

There are two prevalent economic theories of discrimination: *taste-based* and *belief-based*, which arise due to pure preferences [Becker, 1957] and imperfect information, respectively. The latter theory can be particularly informative for the design of fair ML systems as the true attribute of an agent is often not observed directly, but only through a proxy. From this theory, *statistical discrimination* [Arrow, 1973; Phelps, 1972] generally assumes that differences are exogenous but exist. Some papers attribute discrimination to *coordination failure*: agents are born unqualified but can undertake some costly skill investment, which may lead to asymmetric equilibria [Coate and Loury, 1993]. Thus, such economic models offer useful insights on how to design a system aware of inequality due to differentiated skill learning. Finally, another theory of belief-based discrimination is *mis-specification* [Bohren *et al.*, 2019]. Without being aware of their own bias [Pronin *et al.*, 2002], some decision makers may hold misspecified models of group differences which, in the absence of perfect information, lead to false judgment of an individual's abilities.

Beyond this, utilitarianism and normative economics have been used in mechanism design to motivate the use of utility functions as a synonym for social welfare. Although these two terms are used interchangeably their origin differs: as [Posner, 1983] writes, *utilitarianism* is a philosophical system which holds that "the moral worth of an action, practice institution or law is to be judged by its effect on promoting happiness of society." On the other hand, *normative or welfare economics* holds that "an action is to be judged by its effects in promoting the social welfare." In sharp contrast to ML, where multiple definitions of fairness have been used, weighted social welfare is the most commonly accepted measure of broader equity in MD problems.

3 Past and Future Lessons

In this section, we enumerate several lessons that mechanism design (MD) and machine learning (ML) are able to learn from each other. We denote by $A \to B$ a lesson that has been or can be taught by field A to B.

MD → ML: Tension between fairness and welfare. [Kaplow and Shavell, 2003] are among the first to argue that welfare should be the primary metric for the effectiveness of a social policy. They show that optimizing for fairness instead of welfare can actually cause harm in social decision-making processes. This is later supported by [Ben-Porat *et al.*, 2019; Hu and Chen, 2020; Hossain *et al.*, 2020; Kasy and Abebe, 2020], who show that adding group parity constraints can lead to a decrease in welfare for *every* group.

An important question that arises is whether the common utilitarian view of machine learning is problematic. A common criticism is that it is not clear whose utilities we should maximize and how much weight each individual should receive in the optimization objective. For example, should an algorithm ensure the average utilities of both protected and unprotected groups be the same, or should each group contribute to the total welfare proportionally to its size in the society? Using the lens of welfare economics as well as economic theories of discrimination to assess the equitability of ML systems is crucial for designing truly just systems.

MD → ML: Long-term effects of fairness. Because mechanism design considers outcomes for an entire population of agents, the ML community has started to adopt mechanism design techniques (ranging from large market models to equilibria analysis in games to dynamic models of learning agents) in order to study the effects of ML algorithms on different subpopulations. For example, the decisions made by an algorithm can change the population data over time, requiring any "learning" to be dynamic rather than one-shot.

The economics literature has long studied such effects, but not from a machine learning perspective. Although more work is needed to determine whether and how economic models can help inform the design of ML algorithms, some initial progress has been made. [Liu *et al.*, 2018; Kannan *et al.*, 2019] consider two-step models to understand the possibility of harms caused by fairness constraints and the impossibility of equality, respectively. Both papers are strongly influenced by the classic models of [Phelps, 1972] and [Coate and Loury, 1993], respectively. Similarly, [Hu and Chen, 2018] build upon [Levin, 2009] and study the effect of short-term restrictions on improving long-term fairness in labor markets. Drawing upon the theory of mis-specification, [Monachou and Ashlagi, 2019] study the long-term effects of social bias in online labor markets. Using behavioral dynamics, [Heidari *et al.*, 2019] study the temporal relation between social segregation and unfairness.

MD → ML: Strategic agents. The economist's basic analytic tool is the assumption that people are *rational maximizers* of their utility, and most principles of mechanism design are deductions from this basic assumption. An emerging literature utilizes insights from models with strategic incentives to inform machine learning models and is often concerned with agents who can manipulate their features. For example, [Hu *et al.*, 2019] contextualizes strategic investment in test preparation to falsely boost scores that are used as a proxy to quantify college readiness, and the disparate outcomes emerging from the ability to manipulating inputs into a classifier. [Milli *et al.*, 2019; Kleinberg and Raghavan, 2019] similarly show the effects of strategic agents on a classifier and analyze the trade-offs between the decision maker's utility and the social burden different groups incur from their strategies. Thus, it is important to design ML algorithms for decision making with *awareness* of human incentives.

ML → MD: (Re)defining classic notions of fairness. In contrast to the few notions of fairness studied in MD, there seems to be an inexhaustible list of fairness definitions in ML, stemming from its statistical nature. Although a universal definition may be both undesirable and unfeasible, formal definitions are valuable in at least three ways. First, they allow precise reasoning about the normative design decisions involved in building ML systems. Second, they can make clear the ways in which the spirit of fairness can be violated [Corbett-Davies and Goel, 2018; Dwork *et al.*, 2012]. And finally, a profound lesson of fair ML is that intuitive and desirable ideas about fairness may be in conflict; in particular, it may be impossible to simultaneously satisfy multiple fairness notions [Kleinberg *et al.*, 2017], and fairness can impose a penalty to non-fairness desiderata [Corbett-Davies *et al.*, 2017]. Thus, ML can inform MD through its formal definitions that surface inherent tensions, and confront system builders with the inescapable trade-offs they make.

ML → MD: Group-level diagnosis. [Hossain *et al.*, 2020] argue that group-level notions of utility from fair division often focus on improving individual equitability [Conitzer *et al.*, 2019], missing the community impact of an algorithm. On the other hand, group fairness notions from fair ML shed light on legally protected communities that may be subject to disparate impact of a learning system, both in concept, by adapting and using legally protected groups in algorithmic design, and in methodology, through defining group fairness as a relaxation of individual fairness. Thus, ML-inspired notions of group fairness can be leveraged in MD to understand the impact of different welfare functions on different social groups. Beyond this, by considering groups that have legal precedence in being under-served, machine learning can extend the scope of mechanism design to include *group-level diagnosis*. As [Abebe *et al.*, 2020] points out, computing is particularly well-suited for large-scale diagnosis of social inequality, and thus insights from machine learning can be leveraged in understanding social inequalities in mechanism design as well, from the diverse impact of welfare functions on communities, to the way different sub-populations may strategically react to a central planner.

4 Application Domains

Collaboration between ML and MD is motivated by the application domains that they synergistically develop. From *education* and *labor markets*, to *criminal justice* and *ad auctions*, ML and MD must be understood together in the way they bring new perspectives in fairness.

While ML methods have been first shown to exhibit bias in the judiciary sector [Chouldechova, 2017; Jung *et al.*, 2020; Corbett-Davies and Goel, 2018], we are far from achieving a truly 'just' system. Beyond this, both ML and MD have successfully collaborated in applications for labor markets and ad auctions. This involves combining mechanism design ideas with insights about discrimination from the labor economics literature [Hu and Chen, 2018; Kleinberg and Raghavan, 2018] and better understanding how bias and discrimination manifest in nascent domains like algorithmic hiring [Bogen and Rieke, 2018; Raghavan *et al.*, 2020; Sánchez-Monedero *et al.*, 2020] and the gig economy [Rosenblat *et al.*, 2017; Monachou and Ashlagi, 2019]. Finally, while auctions are a subfield of mechanism design, fairness in online ad auctions has largely been inspired by fair ML. Several studies show that the resulting ad deliveries could lead to unfair distribution of audience groups [Sweeney, 2013; Vermeren, 2015; Ali *et al.*, 2019]. This could be due to discriminatory practices or pre-existing bias of the advertisers [Sweeney, 2013; Vermeren, 2015], or even competitive spillovers among advertisers [Ali *et al.*, 2019]. Other recent works have proposed interventions for fairer ad auctions by using suitable group fairness notions [Celis and Vishnoi, 2019] or by extending classical notions like envy-freeness [Ilvento and Chawla, 2020].

While MD has traditionally studied problems such as school choice, college admissions and affirmative action [Abdulkadiroğlu and Sönmez, 2003; Chade *et al.*, 2014; Abdulkadiroğlu, 2005], ML methods have also recently been applied in establishing conditions in which better demographic representation can be achieved in college admissions [Liu *et al.*, 2020; Kannan *et al.*, 2019]. Recent papers [Roth, 2008; Pathak, 2017; Hitzig, 2018] have also pointed to normative gaps in using economic notions of social welfare in solving these problems, showing that this problem is yet far from being solved.

5 Conclusion

Finally, while social problems cannot be solely tackled with tools ML and MD, bridging these two fields is an important step in establishing a pipeline that is ultimately equitable, by incorporating concerns regarding fairness and inequality. Many questions remain open in this field as more work is needed to create meaningful interventions in sociotechnical systems, with applications in labor, education, healthcare, advertising, finance, and social networks.

Acknowledgements

The authors would like to thank Rediet Abebe and Irene Lo for helpful comments and suggestions. This project has been part of the MD4SG working group on Bias, Discrimination, and Fairness.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grants No. 1650115 (Finocchiaro), 1644869 (Maio), 1650441 (Raghavan) and 1761810 (Stoica). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Stoica acknowledges support from the J.P. Morgan AI research fellowship. Monachou acknowledges support from the Krishnan-Shah Fellowship and the A.G. Leventis Foundation Grant. Patro is supported by a fellowship from Tata Consultancy Services Research.

References

[Abdulkadiroğlu and Sönmez, 2003] Atila Abdulkadiroğlu and Tayfun Sönmez. School choice: A mechanism design approach. *American economic review*, 93(3):729–747, 2003.

[Abdulkadiroğlu, 2005] Atila Abdulkadiroğlu. College admissions with affirmative action. *International Journal of Game Theory*, 33(4):535–549, 2005.

[Abebe and Goldner, 2018] Rediet Abebe and Kira Goldner. Mechanism design for social good. *AI Matters*, 4(3):27–34, 2018.

[Abebe *et al.*, 2020] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G Robinson. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 252–260, 2020.

[Ali *et al.*, 2019] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. Discrimination through optimization: How facebook's ad delivery can lead to biased outcomes. In *CSCW*, 2019.

[Angwin *et al.*, 2016] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23:2016, 2016.

[Arrow, 1973] Kenneth Arrow. The theory of discrimination. *Discrimination in labor markets*, 3(10):3–33, 1973.

[Barocas and Selbst, 2016] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[Becker, 1957] Gary S Becker. The economics of discrimination. *University of Chicago Press*, 1957.

[Ben-Porat *et al.*, 2019] Omer Ben-Porat, Fedor Sandomirskiy, and Moshe Tennenholtz. Protecting the protected group: Circumventing harmful fairness. *arXiv preprint arXiv:1905.10546*, 2019.

[Bogen and Rieke, 2018] Miranda Bogen and Aaron Rieke. Help wanted: An examination of hiring algorithms, equity, and bias, 2018.

[Bohren *et al.*, 2019] J Aislinn Bohren, Alex Imas, and Michael Rosenberg. The dynamics of discrimination: Theory and evidence. *American Economic Review*, 109(10):3395–3436, 2019.

[Buolamwini and Gebru, 2018] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

[Celis and Vishnoi, 2019] Anay Mehrotra Celis, Elisa and Nisheeth Vishnoi. Toward controlling discrimination in online ad auctions. In *International Conference on Machine Learning*, 2019.

[Chade *et al.*, 2014] Hector Chade, Gregory Lewis, and Lones Smith. Student portfolios and the college admissions problem. *Review of Economic Studies*, 81(3):971–1002, 2014.

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

[Coate and Loury, 1993] Stephen Coate and Glenn C Loury. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, pages 1220–1240, 1993.

[Conitzer *et al.*, 2019] Vincent Conitzer, Rupert Freeman, Nisarg Shah, and Jennifer Wortman Vaughan. Group fairness for the allocation of indivisible goods. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1853–1860, 2019.

[Corbett-Davies and Goel, 2018] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[Corbett-Davies *et al.*, 2017] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2017.

[Dwork *et al.*, 2012] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[Dworkin, 2002] Ronald Dworkin. *Sovereign virtue: The theory and practice of equality*. Harvard university press, 2002.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[Heidari *et al.*, 2019] Hoda Heidari, Vedant Nanda, and Krishna P Gummadi. On the long-term impact of algorithmic decision policies: Effort unfairness and feature segregation through social learning. *arXiv preprint arXiv:1903.01209*, 2019.

[Hitzig, 2018] Zoë Hitzig. Bridging the 'normative gap': Mechanism design and social justice. *Available at SSRN 3242882*, 2018.

[Hossain *et al.*, 2020] Safwan Hossain, Andjela Mladenovic, and Nisarg Shah. Designing fairly fair classifiers via economic fairness notions. In *Proceedings of The Web Conference 2020*, pages 1559–1569, 2020.

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

[Hu and Chen, 2020] Lily Hu and Yiling Chen. Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 535–545, 2020.

[Hu *et al.*, 2019] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 259–268, 2019.

[Ilvento and Chawla, 2020] Meena Jagadeesan Ilvento, Christina and Shuchi Chawla. Multi-category fairness in sponsored search auctions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.

[Jung *et al.*, 2020] Christopher Jung, Sampath Kannan, Changhwa Lee, Mallesh M Pai, Aaron Roth, and Rakesh Vohra. Fair prediction with endogenous behavior. *ACM Conference on Economics and Computation 2020*, 2020.

[Kannan *et al.*, 2019] Sampath Kannan, Aaron Roth, and Juba Ziani. Downstream effects of affirmative action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 240–248, 2019.

[Kaplow and Shavell, 2003] Louis Kaplow and Steven Shavell. Fairness versus welfare: notes on the pareto principle, preferences, and distributive justice. *The Journal of Legal Studies*, 32(1):331–362, 2003.

[Kasy and Abebe, 2020] Maximilian Kasy and Rediet Abebe. Fairness, equality, and power in algorithmic decision-making. Technical report, Working paper, 2020.

[Kleinberg and Raghavan, 2018] Jon Kleinberg and Manish Raghavan. Selection problems in the presence of implicit bias. *arXiv preprint arXiv:1801.03533*, 2018.

[Kleinberg and Raghavan, 2019] Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 825–844, 2019.

[Kleinberg *et al.*, 2017] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[Levin, 2009] Jonathan Levin. The dynamics of collective reputation. *The BE Journal of Theoretical Economics*, 9(1), 2009.

[Liu *et al.*, 2018] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. *arXiv preprint arXiv:1803.04383*, 2018.

[Liu *et al.*, 2020] Lydia T Liu, Ashia Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs, and Jennifer Chayes. The disparate equilibria of algorithmic decision making when individuals invest rationally. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 381–391, 2020.

[Mehrabi *et al.*, 2019] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

[Milli *et al.*, 2019] Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 230–239, 2019.

[Monachou and Ashlagi, 2019] Faidra Monachou and Itai Ashlagi. Discrimination in online markets: Effects of social bias on learning from reviews and policy design. In *Advances in Neural Information Processing Systems*, pages 2142–2152, 2019.

[Pathak, 2017] Parag A Pathak. What really matters in designing school choice mechanisms. *Advances in Economics and Econometrics*, 1:176–214, 2017.

[Phelps, 1972] Edmund S Phelps. The statistical theory of racism and sexism. *The american economic review*, 62(4):659–661, 1972.

[Posner, 1983] Richard A Posner. *The economics of justice*. Harvard University Press, 1983.

[Pronin *et al.*, 2002] Emily Pronin, Daniel Y Lin, and Lee Ross. The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, 28(3):369–381, 2002.

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

[Rosenblat *et al.*, 2017] Alex Rosenblat, Karen EC Levy, Solon Barocas, and Tim Hwang. Discriminating tastes: Uber's customer ratings as vehicles for workplace discrimination. *Policy & Internet*, 9(3):256–279, 2017.

[Roth, 2008] Alvin E Roth. Deferred acceptance algorithms: History, theory, practice, and open questions. *international Journal of game Theory*, 36(3-4):537–569, 2008.

[Sánchez-Monedero *et al.*, 2020] Javier Sánchez-Monedero, Lina Dencik, and Lilian Edwards. What does it mean to'solve'the problem of discrimination in hiring? social, technical and legal perspectives from the uk on automated hiring systems. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 458–468, 2020.

[Sen, 1979] Amartya Sen. Equality of what. *The Tanner lecture on human values*, 1, 1979.

[Sweeney, 2013] L. Sweeney. Discrimination in online ad delivery. *Commun. ACM, 56(5):44–54*, 2013.

[Verma and Rubin, 2018] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.

[Vermeren, 2015] I Vermeren. Men vs. women: Who is more active on social media? *Brandwatch*, 2015.