Bridging Machine Learning and Mechanism Design towards Algorithmic Fairness

Jessie Finocchiaro
CU Boulder

Roland Maio
Columbia University

Faidra Monachou
Stanford University

Gourab K Patro
IIT Kharagpur

Manish Raghavan
Cornell University

Ana-Andreea Stoica
Columbia University

Stratis Tsirtsis
Max Planck Institute for Software Systems

March 4, 2021

Abstract

Decision-making systems increasingly orchestrate our world: how to intervene on the algorithmic components to build fair and equitable systems is therefore a question of utmost importance; one that is substantially complicated by the context-dependent nature of fairness and discrimination. Modern decision-making systems that involve allocating resources or information to people (e.g., school choice, advertising) incorporate machine-learned predictions in their pipelines, raising concerns about potential strategic behavior or constrained allocation, concerns usually tackled in the context of mechanism design. Although both machine learning and mechanism design have developed frameworks for addressing issues of fairness and equity, in some complex decision-making systems, neither framework is individually sufficient. In this paper, we develop the position that building fair decision-making systems requires overcoming these limitations which, we argue, are inherent to each field. Our ultimate objective is to build an encompassing framework that cohesively bridges the individual frameworks of mechanism design and machine learning. We begin to lay the ground work towards this goal by comparing the perspective each discipline takes on fair decision-making, teasing out the lessons each field has taught and can teach the other, and highlighting application domains that require a strong collaboration between these disciplines.

1 Introduction

Centralized decision-making systems are being increasingly automated through the use of algorithmic tools: user data is processed through algorithms that predict what products and ads a user will click on, student data is used to predict academic performance for admissions into schools and universities, potential employees are increasingly being filtered through algorithms that process their resume data, and so on. Many of these applications have traditionally fallen under the umbrella of mechanism design, from auction design to fair allocation and school matching to labor markets and online platform design. However, recent pushes towards data-driven decision-making have brought together the fields of mechanism design (MD) and machine learning (ML), creating complex pipelines that mediate access to resources and opportunities. Increasingly, learning algorithms are used in the context of mechanism design applications by adopting reinforcement learning techniques in auctions [Feng et al., 2018, Dütting et al., 2019, Zheng et al., 2020, Tang, 2017] or general machine learning algorithms in combinatorial optimization [Bengio et al., 2020] and transportation systems [Jones et al., 2018]. As such applications do not directly focus on fairness and discrimination, they are not the central focus of this paper.

The growing impact of these decision-making and resource allocation systems has prompted an inquiry by computer scientists and economists: are these systems fair and equitable, or do they reproduce or amplify discrimination patterns from our society? In building fair and equitable systems, the question of fairness and discrimination is often a contested one. Paraphrasing Dworkin [2002], *"People who praise or disparage [fairness] disagree about what they are praising or disparaging."* The causes of these philosophical debates include divergent value systems and the context-dependent nature of fairness and discrimination. However, even when we do agree on the types of harms and discrimination we seek to prevent, mechanism design and machine learning often provide different sets of techniques and methodologies to investigate and mitigate these harms. A key goal of this work is to identify the gaps between how machine learning and mechanism design reason about how to treat individuals fairly and detail concrete lessons each field can learn from the other. Our hope is that these lessons will enable more comprehensive analyses of joint ML–MD systems.

Where do the gaps between machine learning and mechanism design come from? Crucially, each field tends to make assumptions or abstractions that can limit the extent to which these interventions perform as desired in practice. This limitation is not specific to machine learning and mechanism design; in general, any field must choose an appropriate scope in which to operate, i.e., a *reducibility assumption*: it is assumed that the issue at hand is reducible to a standard domain problem, and that if the solution to this problem is fair and equitable, then so too will be the overall sociotechnical system [Selbst et al., 2019]. Under the reducibility assumption, fairness and discrimination can be addressed by an intervention that operates within the frame of the field in question, whether that be a constraint on a machine learning algorithm or a balance between the utilities of various agents in a mechanism. Yet, in practice, complex algorithmic decision-making systems rarely satisfy any sort of reducibility assumption; not only do these systems require the combination of ideas from both disciplines, they also depend heavily on the social and cultural contexts in which they operate.

Our goal here is not to argue that it is *sufficient* to consider machine learning and mechanism design in conjunction with one another; rather, we argue that it is *necessary* to do so. Working within each field in isolation will ultimately lead to gaps in our broader understanding of the decision-making systems within which they operate, making it impossible to fully assess the impact these systems have on society. Of course, broadly construed sociotechnical systems cannot be fully understood just through these technical disciplines; our hope is that a more robust understanding of the strengths and weaknesses of machine learning and mechanism design will allow for a clearer view into how they can be integrated into a broader study of these sociotechnical systems.

As an illustrative example, consider the problem of online advertising. Most modern online advertising systems perform a combination of prediction tasks (e.g., how likely is a user to click on this ad?) and allocation tasks (e.g., who should see which ad?). Moreover, these advertising systems significantly impact individuals' lives, including their access to economic opportunity, information, and housing, and new products and technologies (see, e.g., Tobin [2019], Dreyfuss [2019]). Thus, advertising platforms must consider the social impact of their design choices and actively ensure that users are treated fairly.

In isolation, techniques to ensure fair ad distribution from either machine learning or mechanism design fail to fully capture the complexity of the system. On the mechanism design side, auctions typically take learned predictions as a given; as a result, they can overlook the fact that algorithmic predictions are trained on past behavior, which may include the biased bidding and targeting decisions of advertisers. On the other hand, while evaluation tools from fair machine learning would help to ensure that the predictions of interest are "good" for everyone (by some definition), they may fail to capture the externalities of competition between ads that might lead to outcome disparities [Ali et al., 2019]. For example, a job ad may be shown at a higher rate for men than for women because it must compete against a different set of ads targeted at women than at men. As each field has only a partial view of the overall system, it might be impossible to reason about the system's overall impact without taking a broader view that encompasses both the machine learning and mechanism design considerations.

This disconnect is not limited to the ad auction setting described above. Due to their historically different applications and development, both machine learning and mechanism design tend to make different sets of assumptions that do not always hold in practice, especially in pipelines that combine tools from both fields. On the one hand, machine learning traditionally treats people as data points without agency and defines objectives for learning algorithms based on loss functions that depend either on deviations from a ground truth or optimize a pre-defined metric on such data points. Thus, machine learning definitions of fairness tend to ignore complex preferences, long-term effects, and strategic behavior of individuals. On the other hand, as mechanism design often assumes known preferences, and more generally, that information comes from a fixed and known distribution without further critique, and measures utility as a proxy for equality, it tends to miss systematic patterns of discrimination and human perceptions (see also Section 3.5). While recent works have started to address these gaps between machine learning and mechanism design approaches to fairness by embedding welfare notions in measures of accuracy and fairness and using learning algorithms to elicit preferences, many open questions remain on what each field can learn from the other to improve the design of automated decision-making systems.

In this paper, we formalize these ideas into a set of lessons that each field can learn from the other in order to bridge gaps between different theories of fairness and discrimination. In doing so, we aim to provide concrete avenues to address some of the limitations of machine learning and mechanism design, under the acknowledgement that bridging these fields is only an initial step towards a comprehensive analysis of sociotechnical systems.

We make the following contributions:

- We review definitions of fairness and discrimination in machine learning and mechanism design, highlighting historical differences in the way fairness has been defined and implemented in each (Section 2).
- We define several lessons that can be learned from mechanism design and machine learning in order to create an encompassing framework for decision-making. Specifically, we highlight the gap between fairness and welfare, the potential of long-term assessment of decision making systems, group versus individual assessment of fairness and the effect of human perception of fairness, among others (Section 3).
- Finally, we highlight different application domains and survey relevant works in which both mechanism design and machine learning tools have been deployed, such as advertising, education, labor markets and the gig economy, criminal justice, health insurance markets, creditworthiness, and social networks. We discuss advances and limitations of current techniques and implementations in each of these domains, relating to the lessons from the previous section (Section 4).

2 Differences between Mechanism Design and Machine Learning

Machine learning has been increasingly used to supplement human decisions, drawing attention to biases rooted in learning from historically prejudiced data [Angwin et al., 2016, Buolamwini and Gebru, 2018, Barocas and Selbst, 2016]. Fair machine learning often defines fairness conditions (e.g. parity for legally protected groups) without considering core mechanism design concerns such as welfare and strategic behavior. Yet, mechanism design often fails to conceptualize the impact of decisions for different social groups.

While both fields incorporate quantitative notions of fairness into optimization, they differ in the roles those notions play: in machine learning, fairness is typically a constraint to be satisfied, hence the learning algorithms are not optimizing for the *most* fair solution; in contrast, mechanism design typically defines and directly optimizes a fair utility-based objective (e.g., social welfare).

This is only one of many high-level differences between the two fields. Abebe and Goldner [2018] and Kasy and Abebe [2020] indirectly observe that understanding those differences and bridging different notions of fairness is essential in improving access to opportunity for different communities, as well as extending the purpose of each field to encompass the causal effect of algorithmic design on inequality and distribution of power [Kasy and Abebe, 2020].

2.1 Fairness in machine learning

Multiple definitions of fairness have been proposed; interestingly, their common characteristic seems to be that they agree to disagree. Mehrabi et al. [2019] collect the most common fairness definitions; most of them fall into two main categories, *individual* and *group* fairness. Group fairness notions assess the large-scale effect of an algorithmic system on different demographic groups (often defined by legally protected classes). Individual fairness, however, compares outcomes between each individual in a population, requiring people who are similar to each other to receive a similar outcome, and is therefore typically a stronger constraint.

Individual fairness. Inspired by Rawls [2009]' fair equality of opportunity in political philosophy, Dwork et al. [2012] formalize the notion of individual fairness as a constraint in a classification setting where one wants to *"treat similar individuals similarly"* based on a pairwise similarity metric of their features, (partially) designed by domain experts. However, defining similarity metrics is not easy, especially between individuals belonging to sub-populations with different characteristics. Subsequent work, though limited, mainly aims to overcome this obstacle by either learning feature representations that conceal the individuals' membership to a protected group [Zemel et al., 2013, Lahoti et al., 2019] or by selecting individuals based on how they compare in terms of qualification with other members of their own sub-group [Kearns et al., 2017]. However, individual fairness is not equivalent to meritocracy, since qualified covariates might be more difficult to obtain for disadvantaged people, meaning one person may have worked harder to be recognized as "similar" by the algorithm [Hu et al., 2019].

Overall, individual fairness is reminiscent, yet different, from the individual perspective that utility measures in mechanism design often take (e.g., the notion of envy-freeness from mechanism design) and can be used to compare metrics that assess the individual experience in an algorithmic setting. While individual fairness does not take into account one's preferences (often assumed in mechanism design), recent works [Kim et al., 2020] re-design this definition by taking into account individual preferences.

Group fairness. Numerous definitions have been proposed for group fairness [Verma and Rubin, 2018, Mehrabi et al., 2019], suggesting to impose either simple statistical parity conditions between groups [Corbett-Davies et al., 2017] or more complex classification constraints; some aim to equalize each group's opportunity to positive outcomes [Hardt et al., 2016b], balance the misclassification rates among groups [Berk et al., Zafar et al., 2017a] or provide similar classifications under counterfactual group memberships [Kusner et al., 2017]. Kleinberg et al. [2017] and Chouldechova [2017] show that tensions arise when trying to simultaneously achieve multiple notions. However, as Madaio et al. [2020] emphasize, if one is to strive for quantitative fairness, the notion one optimizes for should be context-dependent and developed in partnership with stakeholders.

Despite the variety of individual and group fairness definitions, it becomes apparent that they lack *expressiveness*. Most of these definitions focus solely on the inputs and outputs of the algorithm without taking into account how those outputs ultimately impact real-world outcomes. For example, the most common assumption is that a "positive classification" output is an equally valuable outcome for everyone. As we discuss in Sections 2.2 and 3.1, mechanism design can offer the tools and definitions to overcome such limitations and successfully incorporate important aspects such as individual- and group-level utilities, resource constraints, as well as strategic incentives, to the design of decision-making models.

2.2 Fairness in mechanism design

The mechanism design literature shifts the focus away from *fairness* towards *welfare* and *discrimination*. We review (i) the classic theories of taste-based and statistical discrimination, (ii) utilitarianism and the idealized objective of maximum social welfare, and (iii) fairness in social choice theory.

Economic theories of discrimination. There are two prevalent economic theories of discrimination: *taste-based* and *belief-based*. The key difference between them is the effect of information; taste-based discrimination arises due to pure preferences [Becker, 1957], and persists even with perfect information about individuals. This theory has often been criticized as being simplistic since it is based on the discriminatory principle that decision-making agents derive higher utility from certain social groups [Guryan and Charles, 2013]; however, empirical evidence is rather inconclusive and application-dependent [Charles and Guryan, 2008, Altonji and Pierret, 2001, Knowles et al., 2001, Cui et al., 2017].

The latter theory of belief-based discrimination can be particularly informative for the design of fair machine learning 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. Other 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]. Finally, another belief-based discrimination theory is *mis-specification* [Bohren et al., 2019]; unaware of their own bias [Pronin et al., 2002], 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.

Such economic models offer useful insights on how to design a system aware of inequality due to (i) equilibrium asymmetries, (ii) information limitations, and (iii) human behavioral biases. For example, different social groups may differ in their skill level due to systematic inequalities of opportunity when certain equilibria arise, but not due to inherited differences in their true ability. This may be in sharp contrast to human decision-makers (or even algorithms) who, due to imperfect information or other biases, may incorrectly infer that perceived differences among individuals can be perfectly explained by observed characteristics.

Utilitarianism and normative economics. Beyond discrimination theories, utilitarianism and normative economics have been extensively used in mechanism design to motivate using utility functions as a synonym for social welfare. Although these two terms are used interchangeably and welfare economics is often viewed as applied utilitarianism, 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."* [Posner, 1983] In contrast to machine learning and its multiple definitions of fairness, weighted social welfare is the most accepted measure of broader "social good" in mechanism design but not necessarily of fairness or equity. Typically, utilitarian approaches capture equity by assigning appropriately defined weights to the utility of each agent. Nevertheless, a major limitation remains as welfare economics models rarely explain how to come up with these weights and how to interpret the relative difference between two agents' weights.

Fairness in social choice theory. Social choice theory deals with collective decision making processes, and fairness is of great significance in such processes—particularly in resource allocation problems and voting. In fair allocation, the goal is to divide a resource or set of goods among n agents that is somehow "fair." The literature tends to focus on three primary notions of fairness: *proportional division* [Steihaus, 1948] (every agent receives at least $\frac{1}{n}$ of her perceived value of resources); *equitability* [Foley, 1967] (every agent equally values their allocations); and *envy-freeness* [Varian, 1973] (every agent values their allocation at least as much as another's). While these notions capture fairness of allocations at an individual level, they treat all individuals equally in contrast to individual fairness which relies on some similarity metric to ensure similar outcomes only for similarindividuals. Moreover, in many real-world problems in healthcare, finance, education, relaxed notions of fairness are used due to the hardness of the absolute notions. Conitzer et al. [2019] points out that one deficiency of relaxed notions of fair allocation is that they fail to capture group-level disparities and often leave room for group unfairness (see Section 3.4). Finally, another difference is that, unlike machine learning settings where all individuals prefer positive or higher outcomes, social choice theory can naturally capture different preferences of agents over the possible outcomes.[1]

3 Past and Future Lessons

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.

3.1 MD → ML: Tension between fairness and welfare

Kaplow and Shavell [2003] are among the first to argue, from a legal and economic point of view, that *"the pursuit of notions of fairness results in a needless and, at root, perverse reduction in individuals' well-being,"* and that welfare should be instead the primary metric for the effectiveness of a social policy. Optimizing for fairness instead of welfare can actually cause harm in social decision-making processes (e.g., by leading to a violation of the Pareto principle). This is later supported for quantitative fairness metrics by Hu and Chen [2020], Hossain et al. [2020], who show that adding group parity constraints can decrease welfare for *every* group.

Recent works also propose fairness-to-welfare pathways that transform utility-based metrics into comparing probability of outcome [Zafar et al., 2017b, Balcan et al., 2019, Hossain et al., 2020],

[1] Voting theory deals with aggregating individual preferences. We exclude discussions on voting while we acknowledge the existence of substantial works on fair voting.

showing that fairness definitions do not automatically imply equitable outcomes from a mechanism design perspective, but on the contrary. Kasy and Abebe [2020] formalize some of these tensions, arguing that machine learning definitions fail to acknowledge inequality within protected groups as well as perpetuate it through notions of merit. This is further complicated by the fact that, while notions of fairness in machine learning often treat outcomes as binary with a single desirable outcome, the real world is far more complex; different individuals may have different preferences over a wide range of outcomes. While individual fairness is often incorporating stronger constraints to ensure that individuals receive a good outcome given their features, their preferences are not directly taken into account. Recent works are addressing this gap by re-designing notions of fairness with preferences in mind [Kim et al., 2020].

Using the lens of welfare economics as well as economic theories of discrimination to assess the equitability of machine learning systems can be useful for designing just systems, but it is no panacea. An important question that arises is whether the prevalent utilitarian view of mechanism design is already problematic. A common criticism of utilitarianism 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 society? If we search beyond economics and computer science, we soon realize that practical difficulties and tensions in philosophy, political science, history, sociology and other disciplines are similar to some of the tensions we currently see in machine learning. For example, borrowing from political philosophy, Binns [2018] introduces new notions of fairness that challenge both the common concept of social welfare maximization and fair machine learning definitions, by asking questions such as: *should we minimise the harms to the least advantaged?* In the end, while there may be no universal notion of welfare that adequately captures society's beliefs about whose welfare to prioritize, mechanism design provides the tools to begin to interrogate these welfare trade-offs in a way that machine learning has yet to fully reckon with.

3.2 MD → ML: Long-term effects of fairness

Because mechanism design considers outcomes for an entire population of agents, the machine learning community has started to adopt mechanism design techniques (ranging from equilibria analysis in games to dynamic models of learning agents) in order to study the effects of machine learning algorithms on different subpopulations. For example, the decisions made by an algorithm and the (strategic) participants can change the population data over time, requiring learning to be dynamic rather than one-shot.

Economics has long studied such dynamic effects, but without a machine learning perspective. However, several useful lessons can be extracted from recent works [Zhang and Liu, 2020]. First and foremost, dynamic effects over time are crucial, and, if neglected, they can worsen rather than improve inequality and discrimination in large-scale decision-making systems. Indeed, even simple two-stage models show that it is impossible to achieve full equality and have the potential of causing harm due to fairness constraints [Liu et al., 2018, Kannan et al., 2019]; interestingly, such models and subsequent works [Liu et al., 2020] are strongly influenced by the classic economic models such as Coate and Loury [1993] and Phelps [1972].

Second, the type and complexity of interventions needed to achieve long-term fairness may vary significantly. For example, Hu and Chen [2018] build upon the labor market model in Levin [2009] and showcase the positive effect of simple short-term restrictions (via a group demographic parity constraint) on improving long-term fairness. However, other systems may require a more complex approach; Wen et al. [2019] study fairness in infinite-time dynamics by using a Markov Decision

Process to learn a policy for decision-making that achieves demographic parity or equalized odds in the infinite time dynamics. From a technical perspective, increasing leaning on popular mechanism design tools such as large market models, mean-field equilibria analysis, and dynamic programming techniques seems to be a promising direction for the design of effective and fair policies in machine learning-driven systems.

Finally, most machine learning models focus solely on algorithmic bias and are oblivious to the existence of the social bias that is coming from human agents making complex, dynamic decisions as a response to the system's algorithmic decisions. The interplay between social and algorithmic bias over time may in fact prove itself useful in explaining dynamic patterns of discrimination in sociotechnical systems. Bohren et al. [2019] introduce the discrimination theory of mis-specification and show, both theoretically and empirically, that contradicting patterns of discrimination against women's evaluations in online platforms can be well explained by users' mis-specified bias in sequential ratings. Monachou and Ashlagi [2019] build upon this theory and tools for learning from reviews to study the long-term effects of social bias on worker welfare inequality in online labor markets, while Heidari et al. [2019] also use observational learning to study the temporal relation between social segregation and unfairness.

3.3 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. Therefore, as machine learning algorithms are increasingly used in prescriptive settings, like hiring or loan approval, it becomes necessary to consider the incentives of the agents who are affected from those algorithmic decisions. As transparency laws regarding algorithmic decision-making are gradually being introduced [Voigt and Von dem Bussche, 2017], individuals are now more than ever capable to use insights about the deployed classifiers and accordingly alter their features in order to "game" the system and receive a beneficial outcome.

This observation has initiated a line of work on *strategic classification* [Dalvi et al., 2004, Brückner and Scheffer, 2011, Brückner et al., 2012, Hardt et al., 2016a, Dong et al., 2018, Chen et al., 2020, Hu et al., 2019] which focuses on incentive-aware machine learning algorithms that try to reduce misclassification caused by transparency-induced strategic behavior. The ability to manipulate their features naturally raises several fairness questions. For example, Hu et al. [2019] contextualize strategic investment in test preparation to falsely boost scores that are used as a proxy to quantify college readiness, as well as the disparate equilibria that could potentially emerge in the presence of social groups with disproportionate manipulation capabilities. Additionally, Milli et al. [2019a] utilize credit scoring and lending data to show that there is a trade-off between the utility of a decision-maker who tries to protect themselves from the agents who modify their features strategically and the social burden different groups of agents incur as a consequence.

On a more positive note, recent work has argued that this strategic modification of features does not always correspond to an agent's attempt to "game" the system but could also represent a truthful investment of effort towards improvement, depending on the features being used and the extent to which they can be maliciously manipulated. This idea has become apparent both in the mechanism design literature [Kleinberg and Raghavan, 2019, Alon et al., 2020] on evaluation mechanisms and the machine learning literature [Tsirtsis et al., 2020, Miller et al., 2020, Haghtalab et al., 2020] on the design of transparent decision policies that aim to incentivize the individuals' improvement. Relaxing our initial assumption about strict individual rationality, we can easily see that transparent decision policies based on features prone to manipulation may prove themselves substantially unfair, by equally rewarding seemingly similar individuals with dissimilar effort profiles (in direct opposition to definitions of individual fairness), as those dissimilarities may have ethical, behavioral or cultural origins. For ease of exposition, consider a simple example of admitting graduate students solely based on their undergraduate GPA. Even if two students share the same observable features (GPA), that could reflect different mixtures of manipulating the undergraduate evaluation rules or achieving truthful academic excellence, a behavior often depending on their cultural background [Magnus et al., 2002, Payan et al., 2010]. In this context, the uncertain relation between features and individual qualifications gives rise to a need for *strategyproofness* in order to make prediction-based decision-making systems transparent and fair.

Apart from simple classification settings, the interplay between machine learning and mechanism design also needs to be considered in more complex systems where the stakeholders have more diverse incentives and predictive models of different forms also appear. For example, in health insurance markets machine learning is used to predict the expected costs of individuals and proportionally compensate insurers, with strategic upcoding by the latter favorably skewing subsequent predictions [Cunningham, 2012] and disincentivizing all insurers from offering attractive insurance plans to people with specific medical conditions [Zink and Rose, 2020]. Moreover, the retrieval and recommender systems, well-known downstream applications of machine learning, are also vulnerable to strategic behavior leading to disparate effects even in the absence of model transparency; specifically, strategic manipulation in recommendations [Chakraborty et al., 2019, Song et al., 2020] and search engines [Baruchson-Arbib and Bar-Ilan, 2007, Epstein and Robertson, 2015] often results in skewed information delivery leading to disproportional opportunity or exposure for the users. Such disparate effects of machine learning highlight the need for further research towards the direction of developing models aware of the strategic environment in which they operate as well as the effects of their predictions on different people and groups.

3.4 ML → MD: Defining and diagnosing unfairness under uncertainty

Definitions of fairness from the mechanism design literature tend to be centered around preferences and utilities. As discussed earlier, the fair machine learning literature has yet to fully adopt this perspective, typically operating at the level of model outputs as opposed to the values for individuals produced by those outputs. However, a key assumption necessary for mechanism design's preference-based notions of fairness is that individuals' preferences are known or can be in some way communicated to a central decision-maker. In many mechanism design applications, like traditional auctions or school choice, this assumption can be reasonable. In more complex systems like online advertising, preferences are often unknown a priori and must be estimated in practice. Thus, questions of fairness necessarily involve reasoning about uncertainty and who bears the burden of errors. In this way, ideas about fairness from machine learning can be useful. Because machine learning treats uncertainty as a first class concept, many conceptions of fairness from the machine learning literature explicitly consider errors and their impact on different sub-populations [Hardt et al., 2016b, Chouldechova, 2017, Zafar et al., 2017a].

Uncertainty can also manifest itself with respect to outcomes, not just to preferences. Many application domains utilize probabilistic models—for example, labor market models from mechanism design often consider two-stage processes in which noisy signals provide information about whether a worker is qualified or not [Coate and Loury, 1993, Hu and Chen, 2018]. Importantly, while these models do incorporate uncertainty, the designer knows the true relationship between observed signals and true outcomes, even though this relationship is probabilistic. This style of analysis is less suited to deal with cases where the relationship between signals and ground truth is unknown and can only be learned about through data. The lack of ground-truth information greatly complicates any analysis of the impact of a mechanism, but it is precisely this lack of information that machine learning techniques are designed to handle. Many of the challenges that arise during learning, including data scarcity for certain groups [Buolamwini and Gebru, 2018], feedback loops [Ensign et al., 2018], preference elicitation [Zinkevich et al., 2003, Blum et al., 2004, Goldberg et al., 2020, Frongillo and Waggoner, 2018], and explore-exploit trade-offs [Bird et al., 2016, Immorlica et al., 2019b, Raghavan et al., 2018], implicate serious fairness concerns. By integrating lessons from machine learning on how to define and measure disparities that learning produces, mechanism design can gain a deeper understanding of real-world systems.

Using fairness definitions as a diagnostic tool for potential harms and societal issues is a powerful application of computing, as Abebe et al. [2020] argue. As such, the various group fairness definitions from machine learning focus on illustrating output differences between different legally protected groups, using error measurements to quantify such differences (e.g., false positive/negative rates). A single definition is thus not feasible, nor desirable, but the process of defining fairness has been expanding, both conceptually and practically: from early computer science works that defines fairness through observations [Dwork et al., 2012, Hardt et al., 2016b] or representations [Zemel et al., 2013, Feldman et al., 2015] to understanding causal relationships between features [Kusner et al., 2017, Kilbertus et al., 2017]. While satisfying multiple definitions may not always be possible [Kleinberg et al., 2017], the different definitions of fairness in machine learning offer an opportunity to become more intersectional in defining sensitive groups and in assessing power differentials. More than that, they shift the purpose of defining fairness from a normative one to a diagnostic one, a purpose that mechanism design can learn from when assessing the utility of a system.

Together with a plethora of works from economics that assess differences in welfare at a group level [Coate and Loury, 1993, Hu and Chen, 2018], recent works in mechanism design [Conitzer et al., 2019] propose adapting individual notions of envy-freeness into group-level definitions through stability, e.g., no group of people should prefer the outcome of another group.

The need to assess the outcome differences between groups becomes more pressing as machine learning tools are increasingly being used in traditional mechanism design applications, as previously discussed. Recent works increasingly adapt group fairness methods inspired from machine learning to design fair voting procedures [Celis et al., 2018] and advertising [Kim et al., 2020], bridging the gap between the individual perspective of mechanism design methods and group-level definitions of fairness from machine learning. Beyond transferring lessons from machine learning to mechanism design, we argue that future design must encompass perspectives other than the purely computational one, from sociological understandings of harm and power to economic discrimination and theories of justice.

3.5 ML → MD: Human perceptions and societal expectations of fairness

Early studies on fairness in both mechanism design and machine learning propose various mathematical formulations of fairness, and normatively prescribe how fair decisions should be made. However, given the impossibility to simultaneously satisfy multiple fairness notions [Kleinberg et al., 2017, Chouldechova, 2017], decision-making systems need to be restricted to only selected principles of fairness, a process that becomes challenging in certain applications, such as criminal justice, finance and lending, self-driving cars, and others. Given such applications and their potential for harm, it is essential for the chosen design and principles to be socially acceptable. Thus, there is a need to understand how people assess fairness and how to infer societal expectations about fairness principles in order to account for all voices in a democratic design of decision-making systems.

A line of work [Woodruff et al., 2018, Lee, 2018, Grgic-Hlaca et al., 2018, Green and Chen, 2019, Srivastava et al., 2019, Saha et al., 2020] in machine learning research has taken steps towards this democratization goal through participatory sociotechnical approaches to fairness [Baxter and Sommerville, 2011, Van Dam et al., 2012] by studying human perceptions and societal expectations of fairness. Three major questions emerge from this line of work, which, we argue, are central in developing participatory mechanism design tools that incorporate preferences. We discuss them next.

First, *whose perceptions or assessment of fairness should be considered?* While Awad et al. [2018] and Noothigattu et al. [2018] used crowdsourced preferences from lay humans in the famous moral machine experiment, Jaques [2019] and Yaghini et al. [2019] have argued that preferences should be taken only from relevant individuals (e.g., primary stakeholders, ethicists, domain experts), citing context-dependent aspect of fairness and the possible vulnerability of lay humans to societal biases.

Second, *what options and information should be made available to the participants?* Some studies [Harrison et al., 2020, Saxena et al., 2019, Awad et al., 2018, Noothigattu et al., 2018] directly asked participants to choose the model with the best fairness notion or the best outcomes, whereas others [Srivastava et al., 2019, Grgic-Hlaca et al., 2018, Yaghini et al., 2019] asked indirect questions to infer the acceptable fairness principles (e.g., whether they approve of certain differences in decision outcomes for pairs of individuals from different groups, or the overall outcome distribution). In a different approach, Grgić-Hlača et al. [2018] and Van Berkel et al. [2019] study the validity of using certain input features in the decision-making process in order to achieve procedural fairness.

Finally, *how should the individual preferences be aggregated?* Even though most of the literature has followed some variant of majority rule for this, Noothigattu et al. [2018] and Kahng et al. [2019] have argued for tools like score-based bloc voting or Borda count from voting theory [Elkind et al., 2017] for better representation of participants' choices. These studies have also shown the need of model explainability [Binns et al., 2018, Rader et al., 2018, Dodge et al., 2019], transparency [Rader et al., 2018, Wang et al., 2020], and context-specific feature selection [Grgić-Hlača et al., 2018, Van Berkel et al., 2019] in improving societal fairness perceptions, which mechanism design has traditionally considered as out of scope or assumed to be known, leading to a recent surge in explainability and transparency studies in machine learning. Future work in mechanism design can learn from such studies in challenging current assumptions about preferences, perceptions, and values.

4 Application Domains

In this section, we discuss several application domains of machine learning and mechanism design to illustrate the lessons of Section 3, underscore the complex interplay between these domains, point out gaps, as well as potential ways of bridging these gaps.

We note that many of the applications are open to critique. One might object to the idea of deciding which students are qualified or unqualified to receive an education in college admissions. More fundamentally, one might argue that the overall social system (e.g., criminal justice) in which an application (e.g., recidivism prediction) is embedded is unjust, and further that this cannot be remedied by any technical fairness intervention. We discuss applications merely as an illustration of the lessons we have articulated, and reiterate our position that it is necessary, though not sufficient, to bridge machine learning and mechanism design for algorithmic fairness.

4.1 Online advertising

Auction design (a subfield of mechanism design) deals with the optimal design of allocation and payment rules when a number of agents bid for a resource. As online ad auctions run in a high-frequency online setup that demands automated and precise bidding from the agents, many ad platforms have deployed machine learning models to estimate the relevance of an ad to a customer while using some high-level preferences about advertisers' budget, bidding strategies, and target audiences. Using the automated bids derived from these relevance predictions, ad allocation mechanisms [Ostrovsky and Schwarz, 2011] are run to place specific ads every time a user visits a webpage, thus making the system a complex mix of interdependent components from both machine learning and mechanism design.

Recent studies show that the resulting ad delivery may be problematically skewed; users who differ on sensitive attributes such as gender [Lambrecht and Tucker, 2019], age [Angwin et al., 2017], race [Angwin and Parris Jr, 2016], may receive very different types of ads. For example, search queries with Black-sounding names are highly likely to be shown ads suggestive of arrest records [Sweeney, 2013]. In another study, women were shown relatively fewer advertisements for high-paying jobs than men with similar profiles [Datta et al., 2015]. When ads are about housing, credit or employment, such disparities can harm equality of opportunity.

One cause of problematically skewed ad delivery is explicit targeting of users based on sensitive attributes [Faizullabhoy and Korolova, 2018, Angwin and Parris Jr, 2016], which can be tackled by disallowing ad targeting based on sensitive attributes especially for housing, credit, and employment ads. Although major ad platforms like Google and Facebook had disallowed targeting of opportunities ad based on sensitive attributes, the advertisers could still exploit other personally identifiable information such as area code [Speicher et al., 2018], or using a biased selection of the source audience in the Lookalike audience tool by Facebook. Following a lawsuit [Spinks, 2019], Facebook removed targeting options for housing, credit, and employment ads [Dreyfuss, 2019].

Other studies [Sapiezynski et al., 2019, Ali et al., 2019] again reveal that ad delivery mechanisms could still result in skewed audience distribution based on sensitive attributes even in the absence of any inappropriate targeting. These are often the results of competitive spillovers; relative competition between general opportunity ads and category-specific ads for items like women's fashion can result in opportunity ads being shown to more male audiences. This issue has been tackled from both advertisers' side and auctioneer's side. Solutions on the advertisers' side include running multiple ad campaigns for different sensitive groups (with parity-constrained budgets) [Gelauff et al., 2020], or using different bidding strategies for different demographics groups [Nasr and Tschantz, 2020]. However, such type of targeting has been disallowed by the platforms because of earlier exploitation by discriminatory advertisers. Moreover, rational advertisers may not want to adopt solutions that decrease utility. On the auctioneer's side, the allocation mechanism can be redesigned to ensure fair audience distribution [Dwork and Ilvento, 2019, Ilvento et al., 2020, Chawla and Jagadeesan, 2020, Celis et al., 2019b]. Along with the welfare optimization goal, group fairness constraints can be used to ensure fair audience distribution [Celis et al., 2019b], and individual fairness [Chawla and Jagadeesan, 2020] or envy-freeness constraints [Ilvento et al., 2020] can be adopted to ensure similar individual satisfaction of the users.

Most of these papers have focused on the mechanism design of online ad delivery. Yet all components—advertisers' strategies, platform's relevance prediction, ad allocation mechanism—may be responsible for unfair ad delivery. While the mechanism design components take the relevance predictions from machine learning models as inputs, they often overlook the possibility of biases in these predictions. Thus, to build a fair online ad ecosystem, there is a need to study the role of relevance prediction models and their role in the mechanism design pipeline. In this regard, a line of work in machine learning that studies preference elicitation in auction settings [Parkes, 2005, Zinkevich et al., 2003, Lahaie and Parkes, 2004] can be explored and extended to online advertisements.

4.2 Admissions in education

Schools and universities increasingly use machine learning to inform admissions decisions [Mallett, 2014]. Mechanism design has traditionally studied problems such as school choice, college admissions and affirmative action (e.g., [Abdulkadiroğlu and Sönmez, 2003, Chade et al., 2014, Abdulkadiroğlu, 2005, Chan and Eyster, 2003, Fu, 2006, Kamada and Kojima, 2019, Foster and Vohra, 1992, Immorlica et al., 2019a]). In general, most of these papers adopt similar assumptions and approaches. At their baseline, they model the problem as a two-sided "market" of strategic agents: schools or colleges on the one side and students on the other side. In school choice, the assignment decisions are usually centralized (e.g., all public schools in Boston may commit to a common matching process), while in college admissions, each university decides independently which applicants to admit.

In both cases, explicit fairness considerations are rarely taken into account. The only exception is, of course, *affirmative action*, which is imposed as an additional external constraint on the market. Most economics papers have mostly considered two categories of policies with respect to protected attributes: *group-unaware*

and *group-aware* policies [Fang and Moro, 2011]. Both policy schemes usually translate to demographic parity constraints and similar quota rules.

Interestingly, explicit notions of fairness and equity are less commonly considered. This may be due to various reasons. For example, in a decentralized system such as college admissions, it is unclear whether and—most importantly—how to optimize social welfare. But even in more centralized applications, such as school choice, several dilemmas arise. Given that both market sides have heterogeneous preferences and strategic incentives, should the central planner prioritize the students' or schools' welfare? How is social welfare even practically defined in this case? Indeed, several papers [Roth, 2008, Pathak, 2017, Robertson and Salehi, 2020, Hitzig, Forthcoming, Li, 2017] have offered a broader critique of the approaches used by market designers, pointing to the gap between translating theoretical assumptions to practical solutions.

Machine learning algorithms are increasingly being used in this area as well, for the purpose of parsing data at a large scale more efficiently and embedding missing notions of fairness. The machine learning literature [Haghtalab et al., 2020, Liu et al., 2020, Hu et al., 2019, Emelianov et al., 2020, Immorlica et al., 2019a, Garg et al., 2020] usually poses the admissions problem as a classification task to predict whether an applicant is "qualified" or "unqualified" to attend their university based on covariates given in the student's application (standardized test scores, demographic information, etc.). When framed as a machine learning problem, the task at its core is to accept students who are qualified and reject those who are not. However, when one widens the scope of the problem, one soon realizes that universities have finite capacity for accepting students, which creates market competition and thus strategic incentives among schools and applicants. This latter problem is studied through a dual ML-MD lens by Emelianov et al. [2020], who consider admission policies under implicit bias, and show how affirmative action in the form of group-specific admission thresholds can improve diversity and academic merit at a capacity-constrained university.

Finally, Kannan et al. [2019] highlight another interesting dimension in the intersection of mechanism design, machine learning and policy: downstream effects of affirmative action. The paper draws upon the mechanism design literature to explore how the effects of different policy schemes propagate across education and labor when sequential decisions are made by utility-maximizing agents with potentially conflicting goals (universities vs. employers). They show that fairness notions such as equal opportunity and (strong) irrelevance of group membership can be achieved only in the extreme case where the college does not report grades to the employer. Thus, the problem of intersectionality occurs again: in complex decision pipelines where different fairness metrics may be required yet it may be infeasible to satisfy all simultaneously [Kleinberg et al., 2017], the question of what is an acceptable trade-off between utility maximization and various notions of fairness persists.

4.3 Labor markets and gig economy

Discrimination has been a perennial problem in labor markets. Decades of research has shown that hiring decisions are subject to bias against disadvantaged communities [Bertrand and Mullainathan, 2004, Wenneras and Wold, 2001, Quillian et al., 2017]. More recently, techniques from both machine learning and mechanism design have been brought to bear in the labor market, and in particular, the gig economy, leading to a fresh wave of concern that the persistent discrimination found in traditional labor markets will manifest itself in new and unexpected ways. In particular, we focus on two use cases: employee selection and employee evaluation. Both of these use cases blend techniques from machine learning and mechanism design, and, as we will argue, it is impossible to adequately deal with issues of discrimination and bias without drawing upon ideas from both fields.

Emerging data-driven techniques for employee selection have begun to employ techniques from machine learning to evaluate and sort candidates [Bogen and Rieke, 2018, Raghavan et al., 2020, Sajjadiani et al., 2019]. While some contend that quantitative tools might help to reduce discrimination [Cowgill, 2018], others warn that hiring discrimination will not be solved by machine learning alone [Gosh, 2017]. However, hiring cannot be treated as a purely predictive problem; it requires consideration of factors like allocation, incentives, externalities, and competition, all of which feature more prominently in the mechanism design literature. Consider, for example, the case of salary prediction [Chen et al., 2018]: platforms like LinkedIn use machine learning techniques to predict a job's salary. While this might appear to be a straightforward application of machine learning, it creates strategic incentives that may produce unintended consequences. If a candidate applies to a new position, their potential employer may be able to infer their current salary based on these predictions, enabling the new employer to reduce the salary they offer. Similar consequences can arise from efforts to predict a candidate's likelihood to leave a job [Jayaratne and Jayatilleke, 2020, Sajjadiani et al., 2019]. Moreover, many predictions about candidates are ultimately used in contexts where there is a limited hiring capacity. As a result, predictions about candidates are often later used to rank or filter candidates—a type of mechanism. To avoid the explicit consideration of demographic characteristics, efforts to ensure that candidates are treated fairly (usually through constraints similar to demographic parity) often come at the prediction stage [Raghavan et al., 2020], but fail to make guarantees about the eventual outcomes produced by downstream mechanisms. A more complete effort to prevent discrimination in algorithmic hiring pipelines must leverage the flexibility provided by machine learning to implement anti-discrimination solutions while taking into account the effects of downstream hiring mechanisms.

Beyond issues of discrimination in hiring, recent technological developments have fundamentally changed how labor markets work, particularly with regards to the gig economy, and thus led to a plethora of recent works in this space. For example, Rosenblat et al. [2017] and Monachou and Ashlagi [2019] describe how mechanisms that use customer ratings to evaluate workers can internalize customers' discriminatory tastes. Barzilay and Ben-David [2016] call attention to the ways in which platform design can be used to create or reduce wage disparities. Similarly, Hannák et al. [2017] document the existence of linguistic and other biases in employers' reviews for gig workers on two online labor platforms, TaskRabbit and Fiverr, and the negative effects of gender and racial bias on the number of reviews, rating, search and ranking. Edelman et al. [2017] find evidence of discrimination against African-American guests on Airbnb, highlighting the role that Airbnb's design choices play in facilitating this discrimination. Spurred in part by this work, Airbnb recently launched an initiative to study racial discrimination on their platform [Basu et al., 2020]. Crucially, this body of work combines insights from economics, mechanism design, and machine learning to better understand how discrimination can manifest in the gig economy.

4.4 Criminal justice

Recent popularity of the use of machine learning techniques in prescriptive settings has motivated several attempts to analyze the fairness aspects of predictive and statistical models, especially in the context of a critical application domain like criminal justice. Unsurprisingly, relying on such models in practice can end up reinforcing underlying racial biases, as it has been shown in studies about neighbourhood surveillance [The AI Now Institute, 2019] and recidivism prediction [Angwin et al., 2016]. The latter ProPublica study has raised a heated discussion leading many to advocate that the deployed system, independent of the larger criminal justice system in which it is situated, is plainly unfair. While that was apparent in this instance, a rigorous explanation was not trivial; several responses argued that their claims of discrimination were mainly caused by differences in methodology, like the statistical measure of discrimination [Dieterich et al., 2016, Flores et al., 2016].

Since the deployment of predictive models in the criminal justice system is a contested idea [Harcourt, 2008], knowledge about their potential advantages and pitfalls regarding fairness is crucial in order to perform a fruitful debate on their applicability. As already mentioned in Section 2.1, the proposed theoretical notions of fairness seem to present significant trade-offs [Jung et al., 2020, Corbett-Davies and Goel, 2018] while some of them are impossible to simultaneously satisfy [Chouldechova, 2017, Kleinberg et al., 2017]. Those contradictions naturally raise a major question regarding the criminal justice system and the automated decision-making systems within it: *What do people consider truly fair?* Since there doesn't seem to be a one-size-fits-all answer to this question, a natural step forward is a more participatory approach to the definition of (context-dependent) notions of fairness. As discussed in Section 3.5, some first approaches have been made [Woodruff et al., 2018, Lee, 2018, Grgic-Hlaca et al., 2018, Green and Chen, 2019, Srivastava et al., 2019, Saha et al., 2020] towards studying human perceptions of fairness but questions regarding who are the relevant stakeholders in criminal justice, what notions are more appropriate in that field and how to aggregate preferences still need to be answered. But even under a "perfect" fairness definition, humans involved in the judicial decision-making process might be inherently biased. In this context, machine learning can be leveraged to mitigate these human biases [Valera et al., 2018] and mechanism design can be proven useful in studying the welfare implications and effects on inequality of decisions in criminal justice.

Moreover, merely focusing on the task of fair recidivism prediction might be considered an oversimplification because the assessment of a ML system regarding innocence and guilt ignores both human incentives in the criminal justice pipeline and the humanity of the criminal justice system as a whole. In the United States, a defendant only needs to prove their innocence when their case goes to trial in court. Yet, 95% of felony convictions in the United States are obtained through guilty pleas [gui], and 18% of known exonerees pleaded guilty or did not contest to crimes they did not commit. Machine learning techniques could be applied in conjunction with the critical perspective of mechanism design to better comprehend the racial disparities in both sentence and charge bargaining, as documented by Berdejó [2018]. It is worth noting that any theoretical techniques used to examine the criminal justice system should be wary of the common mechanism design assumption that people are rational expected utility maximizing agents, while desperation, selflessness, or fear often counter this assumption in the real world.

Though enlightening, theoretical understanding of fairness in risk assessment and its aforementioned aspects is not sufficient to suggest adopting the use of such systems. The ultimate decision should be made by the respective stakeholders, considering the practical issues that need to be addressed [Koepke and Robinson, 2018] and the particular context in which risk assessment tools are utilized [Stevenson, 2018].

4.5 Health insurance markets

Interactions between machine learning and mechanism design are salient for fairness in healthcare. For example, prior work studied how machine learning formulations may underpredict black patients' health care needs [Obermeyer et al., 2019]. Here, we draw attention to problems at the intersection of the two fields in health insurance markets.

The Patient Protection and Affordable Care Act (ACA) [ACA, 2010] was designed in part to defuse health-insurer incentives to refuse or avoid coverage to individuals with higher healthcare costs (i.e., selection incentives) [Cunningham, 2012]. One way the ACA addresses this is through risk-adjustment based transfer payments: premiums are transferred from plans with lower expected costs to plans with higher expected costs, compensating insurers and fairly spreading costs. Thus issues of fairness in mechanism design inhere at the policy-design level.

A key component of risk adjustment is estimating individual actuarial risk: inaccurate estimates can create selection incentives for insurers. The Centers for Medicare & Medicaid Services' Hierarchical Condition Category (HCC) model is a widely-used risk adjustment model [cms, 2018, Zink and Rose, 2020]. The HCC predicts an enrollee's expected costs using demographic and diagnosis information; the HCC is therefore group-aware for some protected classes (e.g., sex, age), but is otherwise group-unaware, particularly to groups of enrollees with specific healthcare patterns related (but not limited) to diagnoses, treatments, and prescription-drug use. Although there is evidence that the HCC accurately predicts expected costs for many groups of enrollees, there is also evidence that the HCC makes systematic errors for some groups and that insurers often engage in benefit design to exploit the resulting selection incentives [Jacobs and Sommers, 2015, Geruso et al., 2019].

Thus, healthcare-policy designers, approaching the problem from a mechanism design perspective, encounter the lesson from fair machine learning that it is in general necessary to be group-aware. ML practitioners also encounter the lesson from mechanism design that it is necessary to take into account the strategic behavior of stakeholders. A natural machine-learning oriented response to the systematic error observed in the HCC could be to incorporate more information about enrollees into the model, but because the HCC data are provided by the health insurers, there are concerns that insurers or healthcare providers might then strategically upcode enrollees to favorably skew subsequent predictions [Cunningham, 2012].

Recent work seeks to address these issues in risk adjustment by incorporating fairness interventions to learn a regression model that equalizes systematic error across groups. The proposed fair regression models can bring average predicted costs significantly more in line with average historical costs without a commensurately large penalty to the traditional evaluation metric of R^2 [Zink and Rose, 2020].

We see each discipline's techniques applied in a component-wise fashion towards a competitive health-insurance market that achieves socially optimal outcomes. Notably, neither discipline can independently achieve this goal: selection incentives cannot be defused without accurate risk adjustment; behavior cannot be changed by predictions without appropriately designed incentives.

4.6 Determining creditworthiness

The Financial Technology (FinTech) industry increasingly decides to whom a (home, business) loan should be awarded, and at what interest rate. When one considers banks have finite liquid assets, determining an individual's creditworthiness quickly becomes one piece of a larger problem. Saunders [2019] notes that FinTech can help streamline the application process for loans, among other benefits. However, concerns including disparate impacts of disadvantaged communities, overcharging the poor, the unintelligibility of such algorithms, and protection under consumer laws emerge from the use of machine learning. Overcharging the poor particularly appears to be in part a corollary of determining creditworthiness as a machine learning problem in isolation from mechanism design.

In determining creditworthiness, the true label (ability to pay back the loan) faces two shortcomings: first, it is only observed if the loan is given, and second, it might be a function of the given interest rate. The Pew Research Center [Pew Research Center, 2017] revealed 27.4% of Black applicants and 19.2% of Hispanic applicants were denied mortgages, compared with about 11% of White and Asian applicants. Moreover, when granted a loan, 39% of Black applicants were charged an interest rate over 5%, compared to only 28% of White applicants. This in turn makes repaying the loan more difficult, exacerbating financial insecurities resulting from historical financial and housing oppression, such as loan denial and redlining of neighborhoods. Resulting from the historical imbalance of loan acceptance [Rambachan and Roth, 2019, Liu et al., 2018], Kallus and Zhou [2018] observe that algorithms might still yield "bias in, bias out" phenomena, even with fairness constraints, and online machine learning approaches aim to face this issue by incentivizing exploration [Joseph et al., 2016a,b, Raghavan et al., 2018, Celis et al., 2019a, Schumann et al., 2019, Kilbertus et al., 2020].

As transparency increasingly becomes a legal obligation of financial institutions, such technology is particularly susceptible to disparities [Bartlett et al., 2019], largely because of two tensions shaped by the incentives of different stakeholders. First, individuals who gain insight about the institutions' decision-making processes, might have disproportionate recourse abilities, based on their current financial status and access to opportunity. Since financial institutions are typically for-profit organizations aiming to maximize their utility, Milli et al. [2019a] note that a lack of strategyproofness can disproportionately harm disadvantaged groups in the population. The second tension arises from the fact that financial institutions are asked to find a balance between transparency towards customers and protection of their intellectual property [Milli et al., 2019b, Barocas et al., 2020]. Tsirtsis and Gomez-Rodriguez [2020] argue that maximizing utility in this context may provide limited recourse to disadvantaged populations and they propose methods to counteract such disparities. Those tensions reinforce the need to incorporate the study of incentives in automated decision-making systems before they can be effectively used in financial environments. As Saunders [2019] concludes their report: *The key to FinTech is: Understand first. Proceed with caution.*

4.7 Social networks

Social networks have received scrutiny in the way they reinforce patterns of social inequality and discrimination [McPherson et al., 2001, Calvo-Armengol and Jackson, 2004, DiMaggio and Garip, 2011, Gündoğdu et al., 2019, Okafor, 2020]. Inequality at the level of individual connections is often reinforced by algorithms that use these connections for learning: in opinion diffusion [Fish et al., 2019, Ali et al., 2019, Stoica et al., 2020], recommendation [Stoica et al., 2018], clustering [Chierichetti et al., 2017, Kleindessner et al., 2019], and others. Often, such inequality arises from the individual preference for establishing new connections as well as from pre-defined communities [Avin et al., 2015]. Recent papers discuss these patterns through the lens of welfare economics or equilibrium strategies, with Avin et al. [2018] analyzing the utility function for which preferential attachment is the unique equilibrium solution in a social network. Thus, understanding the incentives behind network creation patterns is crucial for designing better algorithms that learn from relational data and tackling bias at its root cause, as Section 3.2 teaches us.

Beyond this, several works argue that ranking and retrieval algorithms not only reinforce existing bias, but also cause changes in people's behavior [O'Neil, 2016]. To tackle this, a recent line of work takes into consideration the post-ranking and post-recommendation effects in a game-theoretical framework, considering users as players and assigning highly ranked/recommended items to a high pay-off. The lesson from Section 3.3 of modeling individuals as rational agents has started a whole subfield in recommendation systems, starting with Bahar et al. [2016], who focus on finding stable equilibria for which users get the best pay-off for their desired items. Ben-Porat and Tennenholtz [2018] propose new methods, such as the Shapley mediator, to fulfill both fairness and stability conditions (as defined by mechanism design) in cases where content providers are strategic to maximize utility and assume a rational behavior of their users based on their preferences. To account for the incentives of users in post-recommendation settings, Basat et al. [2017], Ben Basat et al. [2015] account for users attempting to promote their own content in information retrieval, describing it as an 'adversarial setting'. The main results point to an increase in general utility when accounting for such incentives, as non-strategic design presents limitations in truly fulfilling individual preferences. Mladenov et al. [2020] directly tackle the problem of welfare by considering recommendations as a resource to be allocated. Incorporating the preferences of the users of a social network in a fair way is thus a subsequent question. Recent works [Chakraborty et al., 2019] tackle this by adapting tools from social choice theory, specifically, by proposing a voting mechanism called Single Transferable Vote to aggregate inferred preferences (votes) of users and achieve better recommendations. This kind of tools can be used to operate in adversarial settings, for example in non-personalised recommendation systems like Twitter or Youtube trending topics, which can be manipulated by flooding the network with bot-created content that can become viral. Methods from mechanism design can be used to protect against strategic behavior that could game the underlying machine learning system, as well as incorporate individual preferences in a meaningful way.

5 Conclusion

While the literature is rapidly growing, many open questions at the intersection of mechanism design and machine learning remain, motivating the need for developing a *lingua franca* of fairness, identifying knowledge gaps and lessons, and ultimately bridging the two fields to work towards a fair pipeline in decision making.

However, both communities must acknowledge that making the pipeline "fair" from a technical perspective does not mean the system is *ipso facto* perfect or just. More interdisciplinary work is needed beyond mechanism design and machine learning to create interventions that improve access to sociotechnical systems and design algorithms for critical application domains.

Acknowledgments. The authors would like to thank Rediet Abebe, Itai Ashlagi, Rafael Frongillo, Nikhil Garg, Jon Kleinberg, Hannah Li, Irene Lo, Francisco Marmolejo-Cossio, Angela Zhou, and anonymous reviewers of the ACM FAccT conference 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. Tsirtsis acknowledges support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 945719).

References

Guilty Plea Problem. *The Innocence Project.* URL `www.guiltypleaproblem.org`.

Public Law 111 - 148: Patient Protection and Affordable Care Act. `https://www.govinfo.gov/app/details/PLAW-111publ148/`, 2010.

Risk adjustment. `https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/Risk-Adjustors`, 2018.

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

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

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

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 ACM Conference on Fairness, Accountability, and Transparency*, pages 252–260, 2020.

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. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019. doi: 10.1145/3359301. URL `https://doi.org/10.1145/3359301`.

Tal Alon, Magdalen Dobson, Ariel D Procaccia, Inbal Talgam-Cohen, and Jamie Tucker-Foltz. Multiagent evaluation mechanisms. In *Association for the Advancement of Artificial Intelligence*, pages 1774–1781, 2020.

Joseph G Altonji and Charles R Pierret. Employer learning and statistical discrimination. *The quarterly journal of economics*, 116(1):313–350, 2001.

Julia Angwin and Terry Parris Jr. Facebook lets advertisers exclude users by race. *ProPublica blog*, 28, 2016.

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

Julia Angwin, Noam Scheiber, and Ariana Tobin. Facebook job ads raise concerns about age discrimination. *The New York Times*, 20:1, 2017.

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

Chen Avin, Barbara Keller, Zvi Lotker, Claire Mathieu, David Peleg, and Yvonne-Anne Pignolet. Homophily and the glass ceiling effect in social networks. In *Conference on Innovations in Theoretical Computer Science*, pages 41–50, 2015.

Chen Avin, Avi Cohen, Pierre Fraigniaud, Zvi Lotker, and David Peleg. Preferential attachment as a unique equilibrium. In *World Wide Web Conference*, pages 559–568, 2018.

Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

Gal Bahar, Rann Smorodinsky, and Moshe Tennenholtz. Economic Recommendation Systems: One Page Abstract. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, EC '16, page

757, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450339360. doi: 10.1145/2940716.2940719. URL `https://doi.org/10.1145/2940716.2940719`.

Maria-Florina F Balcan, Travis Dick, Ritesh Noothigattu, and Ariel D Procaccia. Envy-free classification. In *Advances in Neural Information Processing Systems*, pages 1238–1248, 2019.

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

Solon Barocas, Andrew D Selbst, and Manish Raghavan. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency*, pages 80–89, 2020.

Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace. Consumer-lending discrimination in the fintech era. Technical report, National Bureau of Economic Research, 2019.

Shifra Baruchson-Arbib and Judit Bar-Ilan. Manipulating search engine algorithms: the case of google. *Journal of Information, Communication and Ethics in Society*, 2007.

Arianne Renan Barzilay and Anat Ben-David. Platform inequality: gender in the gig-economy. *Seton Hall L. Rev.*, 47:393, 2016.

Ran Ben Basat, Moshe Tennenholtz, and Oren Kurland. A game theoretic analysis of the adversarial retrieval setting. *Journal of Artificial Intelligence Research*, 60:1127–1164, 2017.

Sid Basu, Ruthie Berman, Adam Bloomston, John Campbell, Anne Diaz, Nanako Era, Benjamin Evans, Sukhada Palkar, and Skyler Wharton. Measuring discrepancies in airbnb guest acceptance rates using anonymized demographic data. Technical report, Airbnb, 2020.

Gordon Baxter and Ian Sommerville. Socio-technical systems: From design methods to systems engineering. *Interacting with computers*, 23(1):4–17, 2011.

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

Ran Ben Basat, Moshe Tennenholtz, and Oren Kurland. The probability ranking principle is not optimal in adversarial retrieval settings. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, pages 51–60, 2015.

Omer Ben-Porat and Moshe Tennenholtz. A game-theoretic approach to recommendation systems with strategic content providers. In *Advances in Neural Information Processing Systems*, pages 1110–1120, 2018.

Yoshua Bengio, Andrea Lodi, and Antoine Prouvost. Machine learning for combinatorial optimization: a methodological tour d'horizon. *European Journal of Operational Research*, 2020.

Carlos Berdejó. Criminalizing race: Racial disparities in plea-bargaining. *BCL Rev.*, 59:1187, 2018.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*.

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

Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In *Proceedings of the 2018 ACM Conference on Fairness, Accountability and Transparency*, pages 149–159, 2018.

Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 'it's reducing a human being to a percentage' perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2018.

Sarah Bird, Solon Barocas, Kate Crawford, Fernando Diaz, and Hanna Wallach. Exploring or exploiting? Social and ethical implications of autonomous experimentation in AI. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2016.

Avrim Blum, Jeffrey Jackson, Tuomas Sandholm, and Martin Zinkevich. Preference elicitation and query learning. *Journal of Machine Learning Research*, 5(Jun):649–667, 2004.

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

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

Michael Brückner and Tobias Scheffer. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 547–555, 2011.

Michael Brückner, Christian Kanzow, and Tobias Scheffer. Static prediction games for adversarial learning problems. *The Journal of Machine Learning Research*, 13(1):2617–2654, 2012.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 2018 ACM Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

Antoni Calvo-Armengol and Matthew O Jackson. The effects of social networks on employment and inequality. *American economic review*, 94(3):426–454, 2004.

L Elisa Celis, Lingxiao Huang, and Nisheeth K Vishnoi. Multiwinner voting with fairness constraints. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018.

L Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth Vishnoi. Controlling polarization in personalization: An algorithmic framework. In *Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency*, pages 160–169, 2019a.

L Elisa Celis, Anay Mehrotra, and Nisheeth K Vishnoi. Toward controlling discrimination in online ad auctions. In *Proceedings of the 36th International Conference on Machine Learning*, 2019b.

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

Abhijnan Chakraborty, Gourab K Patro, Niloy Ganguly, Krishna P Gummadi, and Patrick Loiseau. Equality of voice: Towards fair representation in crowdsourced top-k recommendations. In *Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency*, pages 129–138, 2019.

Jimmy Chan and Erik Eyster. Does banning affirmative action lower college student quality? *American Economic Review*, 93(3):858–872, 2003.

Kerwin Kofi Charles and Jonathan Guryan. Prejudice and wages: an empirical assessment of becker's the economics of discrimination. *Journal of political economy*, 116(5):773–809, 2008.

Shuchi Chawla and Meena Jagadeesan. Fairness in ad auctions through inverse proportionality. *arXiv preprint arXiv:2003.13966*, 2020.

Xi Chen, Yiqun Liu, Liang Zhang, and Krishnaram Kenthapadi. How LinkedIn economic graph bonds information and product: applications in LinkedIn salary. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 120–129, 2018.

Y Chen, Y Liu, and C Podimata. Learning strategy-aware linear classifiers. *arXiv preprint arXiv:1911.04004*, 2020.

Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. In *Advances in Neural Information Processing Systems*, pages 5029–5037, 2017.

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

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

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.

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.

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.

Bo Cowgill. Bias and productivity in humans and algorithms: Theory and evidence from resume screening. *Columbia Business School, Columbia University*, 29, 2018.

Ruomeng Cui, Jun Li, and Dennis Zhang. Discrimination with incomplete information in the sharing economy: Evidence from field experiments on airbnb. *Harvard Business School*, pages 1–35, 2017.

Rob Cunningham. Risk adjustment in health insurance. *Health Affairs Policy Brief*, 2012.

Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 99–108, 2004.

Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *Proceedings on Privacy Enhancing Technologies*, 2015(1): 92–112, 2015.

William Dieterich, Christina Mendoza, and Tim Brennan. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 2016.

Paul DiMaggio and Filiz Garip. How network externalities can exacerbate intergroup inequality. *American Journal of Sociology*, 116(6):1887–1933, 2011.

Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. Explaining models: An empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 275–285, 2019.

Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70, 2018.

Emily Dreyfuss. Facebook changes its ad tech to stop discrimination. *Wired*, March 2019. URL `https://www.wired.com/story/facebook-advertising-discrimination-settlement/`.

Paul Dütting, Zhe Feng, Harikrishna Narasimhan, David Parkes, and Sai Srivatsa Ravindranath. Optimal auctions through deep learning. In *International Conference on Machine Learning*, pages 1706–1715, 2019.

Cynthia Dwork and Christina Ilvento. Fairness under composition. In *10th Innovations in Theoretical Computer Science*, 2019.

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.

Ronald Dworkin. *Sovereign virtue: The theory and practice of equality*. Harvard University Press, 2002.

Benjamin Edelman, Michael Luca, and Dan Svirsky. Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics*, 9(2):1–22, 2017.

Edith Elkind, Piotr Faliszewski, Piotr Skowron, and Arkadii Slinko. Properties of multiwinner voting rules. *Social Choice and Welfare*, 48(3):599–632, 2017.

Vitalii Emelianov, Nicolas Gast, Krishna P Gummadi, and Patrick Loiseau. On fair selection in the presence of implicit variance. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 649–675, 2020.

Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *Proceedings of the 2018 ACM Conference on Fairness, Accountability and Transparency*, pages 160–171, 2018.

Robert Epstein and Ronald E Robertson. The search engine manipulation effect (seme) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33):E4512–E4521, 2015.

Irfan Faizullabhoy and Aleksandra Korolova. Facebook's advertising platform: New attack vectors and the need for interventions. *arXiv preprint arXiv:1803.10099*, 2018.

Hanming Fang and Andrea Moro. Theories of statistical discrimination and affirmative action: A survey. In *Handbook of social economics*, volume 1, pages 133–200. Elsevier, 2011.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

Zhe Feng, Harikrishna Narasimhan, and David C Parkes. Deep learning for revenue-optimal auctions with budgets. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, pages 354–362, 2018.

Benjamin Fish, Ashkan Bashardoust, Danah Boyd, Sorelle Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. Gaps in Information Access in Social Networks? In *The World Wide Web Conference*, pages 480–490, 2019.

Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation*, 80:38, 2016.

Duncan Karl Foley. Resource allocation and the public sector. 1967.

Dean P Foster and Rakesh V Vohra. An economic argument for affirmative action. *Rationality and Society*, 4(2):176–188, 1992.

Rafael Frongillo and Bo Waggoner. An axiomatic study of scoring rule markets. In *Innovations in Theoretical Computer Science Conference (ITCS)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

Qiang Fu. A theory of affirmative action in college admissions. *Economic Inquiry*, 44(3):420–428, 2006.

Nikhil Garg, Hannah Li, and Faidra Monachou. Standardized tests and affirmative action: The role of bias and variance. *arXiv preprint arXiv:2010.04396*, 2020.

Lodewijk Gelauff, Ashish Goel, Kamesh Munagala, and Sravya Yandamuri. Advertising for demographically fair outcomes. *arXiv preprint arXiv:2006.03983*, 2020.

Michael Geruso, Timothy Layton, and Daniel Prinz. Screening in contract design: Evidence from the ACA health insurance exchanges. *American Economic Journal: Economic Policy*, 11(2):64–107, 2019.

Paul W Goldberg, Edwin Lock, and Francisco Marmolejo-Cossío. Learning strong substitutes demand via queries. *arXiv preprint arXiv:2005.01496*, 2020.

Dipayan Gosh. Ai is the future of hiring, but it's far from immune to bias. *Quartz at Work*, 17, 2017.

Ben Green and Yiling Chen. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pages 90–99, 2019.

Nina Grgic-Hlaca, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference*, pages 903–912, 2018.

Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Didem Gündoğdu, Pietro Panzarasa, Nuria Oliver, and Bruno Lepri. The bridging and bonding structures of place-centric networks: Evidence from a developing country. *PloS one*, 14(9):e0221148, 2019.

Jonathan Guryan and Kerwin Kofi Charles. Taste-based or statistical discrimination: the economics of discrimination returns to its roots. *The Economic Journal*, 123(572):F417–F432, 2013.

Nika Haghtalab, Nicole Immorlica, Brendan Lucier, and Jack Wang. Maximizing welfare with incentive-aware evaluation mechanisms. In *29th International Joint Conference on Artificial Intelligence*, 2020.

Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. Bias in online freelance marketplaces: Evidence from TaskRabbit and Fiverr. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1914–1933, 2017.

Bernard E Harcourt. *Against prediction: Profiling, policing, and punishing in an actuarial age*. University of Chicago Press, 2008.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 111–122, 2016a.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016b.

Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency*, pages 392–402, 2020.

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. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Zoë Hitzig. The Normative Gap: Mechanism Design and Ideal Theories of Justice. *Economics & Philosophy*, Forthcoming.

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.

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.

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

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

Christina Ilvento, Meena Jagadeesan, and Shuchi Chawla. Multi-category fairness in sponsored search auctions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 348–358, 2020.

Nicole Immorlica, Katrina Ligett, and Juba Ziani. Access to population-level signaling as a source of inequality. In *Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency*, pages 249–258, 2019a.

Nicole Immorlica, Jieming Mao, and Christos Tzamos. Diversity and exploration in social learning. In *The World Wide Web Conference*, pages 762–772, 2019b.

Douglas Bernard Jacobs and Benjamin Daniel Sommers. Using drugs to discriminate—adverse selection in the insurance marketplace. *New England Journal of Medicine*, 2015.

Abby Everett Jaques. Why the moral machine is a monster. *University of Miami School of Law*, 10, 2019.

Madhura Jayaratne and Buddhi Jayatilleke. Predicting job-hopping likelihood using answers to open-ended interview questions. *arXiv preprint arXiv:2007.11189*, 2020.

Philip B Jones, Jonathan Levy, Jeniifer Bosco, John Howat, and John W Van Alst. The future of transportation electrification: Utility, industry and consumer perspectives. Technical report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2018.

Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Fair algorithms for infinite and contextual bandits. *arXiv preprint arXiv:1610.09559*, 2016a.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pages 325–333, 2016b.

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.

Anson Kahng, Min Kyung Lee, Ritesh Noothigattu, Ariel Procaccia, and Christos-Alexandros Psomas. Statistical foundations of virtual democracy. In *International Conference on Machine Learning*, pages 3173–3182, 2019.

Nathan Kallus and Angela Zhou. Residual unfairness in fair machine learning from prejudiced data. In *International Conference on Machine Learning*, pages 2439–2448, 2018.

Yuichiro Kamada and Fuhito Kojima. Fair matching under constraints: Theory and applications. Technical report, 2019.

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

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.

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

Michael Kearns, Aaron Roth, and Zhiwei Steven Wu. Meritocratic fairness for cross-population selection. In *International Conference on Machine Learning*, pages 1828–1836, 2017.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in neural information processing systems*, pages 656–666, 2017.

Niki Kilbertus, Manuel Gomez Rodriguez, Bernhard Schölkopf, Krikamol Muandet, and Isabel Valera. Fair decisions despite imperfect predictions. In *International Conference on Artificial Intelligence and Statistics*, pages 277–287, 2020.

Michael P Kim, Aleksandra Korolova, Guy N Rothblum, and Gal Yona. Preference-informed fairness. In *Innovations in Theoretical Computer Science*, 2020.

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.

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.

Matthäus Kleindessner, Pranjal Awasthi, and Jamie Morgenstern. Fair k-center clustering for data summarization. In *International Conference on Machine Learning*, pages 3448–3457, 2019.

John Knowles, Nicola Persico, and Petra Todd. Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy*, 109(1):203–229, 2001.

John Logan Koepke and David G Robinson. Danger ahead: Risk assessment and the future of bail reform. *Wash. L. Rev.*, 93:1725, 2018.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

Sebastien M Lahaie and David C Parkes. Applying learning algorithms to preference elicitation. In *Proceedings of the 5th ACM conference on Electronic commerce*, pages 180–188, 2004.

Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. ifair: Learning individually fair data representations for algorithmic decision making. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1334–1345. IEEE, 2019.

Anja Lambrecht and Catherine Tucker. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science*, 65(7):2966–2981, 2019.

Min Kyung Lee. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1), 2018.

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

Shengwu Li. Ethics and market design. *Oxford Review of Economic Policy*, 33(4):705–720, 2017.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

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 ACM Conference on Fairness, Accountability, and Transparency*, pages 381–391, 2020.

Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

Jan R Magnus, Victor M Polterovich, Dmitri L Danilov, and Alexei V Savvateev. Tolerance of cheating: An analysis across countries. *The Journal of Economic Education*, 33(2):125–135, 2002.

Whitney Mallett. Behind the color-blind diversity algorithm for college admissions. 2014. URL `https://www.vice.com/en/article/nzee5d/behind-the-color-blind-college-admissions-diversity-algorithm`.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

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.

John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

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

Smitha Milli, Ludwig Schmidt, Anca D Dragan, and Moritz Hardt. Model reconstruction from model explanations. In *Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency*, pages 1–9, 2019b.

Martin Mladenov, Elliot Creager, Omer Ben-Porat, Kevin Swersky, Richard Zemel, and Craig Boutilier. Optimizing Long-term Social Welfare in Recommender Systems:A Constrained Matching Approach. In *Proceedings of the Thirty-seventh International Conference on Machine Learning (ICML-20)*, Vienna, Austria, 2020. to appear.

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.

Milad Nasr and Michael Carl Tschantz. Bidding strategies with gender nondiscrimination constraints for online ad auctions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 337–347, 2020.

Ritesh Noothigattu, Snehalkumar S Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D Procaccia. A voting-based system for ethical decision making. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Chika O Okafor. All things equal? social networks as a mechanism for discrimination. *arXiv preprint arXiv:2006.15988*, 2020.

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

Michael Ostrovsky and Michael Schwarz. Reserve prices in internet advertising auctions: A field experiment. In *Proceedings of the 12th ACM Conference on Electronic commerce*, pages 59–60, 2011.

David C Parkes. Auction design with costly preference elicitation. *Annals of Mathematics and Artificial Intelligence*, 44(3):269–302, 2005.

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

Janice Payan, James Reardon, and Denny E McCorkle. The effect of culture on the academic honesty of marketing and business students. *Journal of Marketing Education*, 32(3):275–291, 2010.

Pew Research Center. Blacks, Hispanics more likely to pay higher mortgage rates. *The Pew Research Center*, 2017.

Edmund S Phelps. The statistical theory of racism and sexism. *The American Economic Review*, 62(4): 659–661, 1972.

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

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.

Lincoln Quillian, Devah Pager, Ole Hexel, and Arnfinn H Midtbøen. Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. *Proceedings of the National Academy of Sciences*, 114(41):10870–10875, 2017.

Emilee Rader, Kelley Cotter, and Janghee Cho. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13, 2018.

Manish Raghavan, Aleksandrs Slivkins, Jennifer Wortman Vaughan, and Zhiwei Steven Wu. The externalities of exploration and how data diversity helps exploitation. *arXiv preprint arXiv:1806.00543*, 2018.

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

Ashesh Rambachan and Jonathan Roth. Bias in, bias out? Evaluating the folk wisdom. *arXiv preprint arXiv:1909.08518*, 2019.

John Rawls. *A theory of justice*. Harvard university press, 2009.

Samantha Robertson and Niloufar Salehi. What If I Don't Like Any Of The Choices? The Limits of Preference Elicitation for Participatory Algorithm Design. *arXiv preprint arXiv:2007.06718*, 2020.

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.

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

Debjani Saha, Candice Schumann, Duncan C McElfresh, John P Dickerson, Michelle L Mazurek, and Michael Carl Tschantz. Measuring non-expert comprehension of machine learning fairness metrics. In *International Conference on Machine Learning*, 2020.

Sima Sajjadiani, Aaron J Sojourner, John D Kammeyer-Mueller, and Elton Mykerezi. Using machine learning to translate applicant work history into predictors of performance and turnover. *Journal of Applied Psychology*, 2019.

Piotr Sapiezynski, Avijit Gosh, Levi Kaplan, Alan Mislove, and Aaron Rieke. Algorithms that "Don't See Color": Comparing Biases in Lookalike and Special Ad Audiences. *arXiv preprint arXiv:1912.07579*, 2019.

Lauren Saunders. FinTech and Consumer Protection: A Snapshot. 2019.

Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 99–106, 2019.

Candice Schumann, Zhi Lang, Nicholas Mattei, and John P Dickerson. Group fairness in bandit arm selection. *arXiv preprint arXiv:1912.03802*, 2019.

Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency*, pages 59–68, 2019.

Junshuai Song, Zhao Li, Zehong Hu, Yucheng Wu, Zhenpeng Li, Jian Li, and Jun Gao. PoisonRec: An Adaptive Data Poisoning Framework for Attacking Black-box Recommender Systems. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 157–168. IEEE, 2020.

Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna Gummadi, Patrick Loiseau, and Alan Mislove. Potential for discrimination in online targeted advertising. In *Proceedings of the 2018 ACM Conference on Fairness, Accountability, and Transparency*, volume 81, pages 1–15, 2018.

Chandler Nicholle Spinks. Contemporary Housing Discrimination: Facebook, Targeted Advertising, and the Fair Housing Act. *Hous. L. Rev.*, 57:925, 2019.

Megha Srivastava, Hoda Heidari, and Andreas Krause. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2459–2468, 2019.

Hugo Steihaus. The problem of fair division. *Econometrica*, 16:101–104, 1948.

Megan Stevenson. Assessing risk assessment in action. *Minn. L. Rev.*, 103:303, 2018.

Ana-Andreea Stoica, Christopher Riederer, and Augustin Chaintreau. Algorithmic glass ceiling in social networks: The effects of social recommendations on network diversity. In *Proceedings of the 2018 World Wide Web Conference*, pages 923–932, 2018.

Ana-Andreea Stoica, Jessy Xinyi Han, and Augustin Chaintreau. Seeding network influence in biased networks and the benefits of diversity. In *Proceedings of The Web Conference 2020*, pages 2089–2098, 2020.

Latanya Sweeney. Discrimination in online ad delivery. *Queue*, 11(3):10–29, 2013.

Pingzhong Tang. Reinforcement mechanism design. In *IJCAI*, pages 5146–5150, 2017.

The AI Now Institute. AI Now 2019 report. Technical report, 2019.

Ariana Tobin. Facebook changes its ad tech to stop discrimination. *ProPublica*, March 2019. URL `https://www.propublica.org/article/hud-sues-facebook-housing-discrimination-advertising-algorithms`.

Stratis Tsirtsis and Manuel Gomez-Rodriguez. Decisions, counterfactual explanations and strategic behavior. In *34th Conference on Neural Information Processing Systems*, 2020.

Stratis Tsirtsis, Behzad Tabibian, Moein Khajehnejad, Adish Singla, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. Optimal decision making under strategic behavior. *arXiv preprint arXiv:1905.09239*, 2020.

Isabel Valera, Adish Singla, and Manuel Gomez Rodriguez. Enhancing the accuracy and fairness of human decision making. In *Advances in Neural Information Processing Systems*, pages 1769–1778, 2018.

Niels Van Berkel, Jorge Goncalves, Danula Hettiachchi, Senuri Wijenayake, Ryan M Kelly, and Vassilis Kostakos. Crowdsourcing perceptions of fair predictors for machine learning: a recidivism case study. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–21, 2019.

Koen H Van Dam, Igor Nikolic, and Zofia Lukszo. *Agent-based modelling of socio-technical systems*, volume 9. Springer Science & Business Media, 2012.

Hal R Varian. Equity, envy, and efficiency. 1973.

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

Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.

Ruotong Wang, F Maxwell Harper, and Haiyi Zhu. Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

Min Wen, Osbert Bastani, and Ufuk Topcu. Fairness with dynamics. *arXiv preprint arXiv:1901.08568*, 2019.

Christine Wenneras and Agnes Wold. Nepotism and sexism in peer-review. *Women, Science and Technology: A Reader in Feminist Science Studies*, pages 46–52, 2001.

Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2018.

Mohammad Yaghini, Hoda Heidari, and Andreas Krause. A human-in-the-loop framework to construct context-dependent mathematical formulations of fairness. *arXiv preprint arXiv:1911.03020*, 2019.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017a.

Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, pages 229–239, 2017b.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

Xueru Zhang and Mingyan Liu. Fairness in Learning-Based Sequential Decision Algorithms: A Survey. *arXiv preprint arXiv:2001.04861*, 2020.

Stephan Zheng, Alexander Trott, Sunil Srinivasa, Nikhil Naik, Melvin Gruesbeck, David C Parkes, and Richard Socher. The ai economist: Improving equality and productivity with ai-driven tax policies. *arXiv preprint arXiv:2004.13332*, 2020.

Anna Zink and Sherri Rose. Fair regression for health care spending. *Biometrics*, 76(3):973–982, 2020.

Martin A Zinkevich, Avrim Blum, and Tuomas Sandholm. On polynomial-time preference elicitation with value queries. In *Proceedings of the 4th ACM Conference on Electronic Commerce*, pages 176–185, 2003.