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

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 systems incorporate machine-learned predictions in broader decision-making pipelines, implicating concerns like constrained allocation and strategic behavior that are typically thought of as mechanism design problems. Although both machine learning and mechanism design have individually 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 the individual frameworks of machine learning and mechanism design. Our ultimate objective is to build an encompassing framework that cohesively bridges the individual frameworks. We begin to lay the ground work towards achieving this goal by comparing the perspective each individual 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 [68, 80, 185, 203] or general machine learning algorithms in combinatorial optimization [31] and transportation systems [114]. 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-allocating 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 [71], *"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 [174]. 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 mechanism design and machine learning, 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. Taking each field in isolation will ultimately lead to gaps in our broader understanding of the decision-making systems in which they operate, making it impossible to fully assess the impacts these systems have on society. Of course, sociotechnical systems broadly construed 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 the impact of 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.[1] Thus, advertising platforms must consider the social impacts 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 [9]. 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 and measures utility as a proxy for equality, it tends to miss systematic patterns of discrimination and human perceptions. 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

Over time, machine learning has been increasingly used to supplement human decisions, drawing attention to biases rooted in learning from historically prejudiced data [12, 20, 44]. Fair machine learning often relies on establishing parity conditions for legally protected groups without considering concepts at the core of mechanism design, such as welfare notions or strategic behavior. However, mechanism design often fails to conceptualize the impact of decisions for different social groups. While both fields approach fairness and discrimination through an engineering lens, they differ in definitions. For example, given machine learning's focus on quantitative fairness definitions, we observe that fair machine learning algorithms are not *optimizing* for the most fair solution; rather, fairness is often seen as a constraint on the set of feasible solutions. This is in contrast to the typical approach in mechanism design which involves directly maximizing utility-based objectives, like social welfare.

Of course, this is only one of many high-level differences between the two fields. Abebe and Goldner [7] and Kasy and Abebe [121] indirectly observe that understanding those differences and

[1] Take, for example, the recent *HUD vs. Facebook* case [67, 187].

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 [121].

2.1 Fairness in machine learning

Multiple quantitative definitions of fair machine learning algorithms have been proposed; interestingly, their common characteristic seems to be that they agree to disagree. Mehrabi et al. [142] have collected the most common fairness definitions (Appendix A.1); most of them fall into two main categories, *individual* and *group* fairness. Individual fairness imposes a much stronger constraint compared to group fairness on what constitutes a fair treatment: an algorithm must be fair with respect to every single individual. On the other hand, in group fairness definitions, an algorithm strives for the average treatment of members of different groups, usually distinguished by a feature like race or gender, to be equal.

Individual fairness. Inspired by work on fair equality of opportunity by Rawls [165] in political philosophy, Dwork et al. [70] 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, such similarity metrics are often not easy to design, especially between individuals belonging to different protected groups. Moreover, individual fairness is not necessarily the golden standard, since qualified covariates might be more difficult to obtain for disadvantaged people, meaning one person may have overcome much more and worked harder to be recognized as "similar" by the algorithm [107].

To overcome these limitations, Zemel et al. [201] argued that finding a proper similarity metric which leads to individual fairness can be reduced to a representation learning problem, aiming to map individuals in a new feature space preserving as much information as possible about their initial features while concealing any information related to their membership in a protected group. In a slightly different and more recent line of work, notions of individual fairness have also come up in the context of rankings with Biega et al. [35] proposing a concept of *amortized individual fairness*, in settings where ranked individuals of similar quality receive dissimilar levels of attention caused by the ranking's position bias.

Group fairness. In contrast to individual fairness, *group fairness* notions assess the bias of a system's treatment of different groups (often defined by legally protected classes) through notions such as equalized odds, equal opportunity, demographic parity, treatment equality, and conditional statistical parity [100, 142, 193], often defined as a relaxed version of individual fairness and aiming to assess the large-scale effect of an algorithm on vulnerable populations. Kleinberg et al. [125] and Chouldechova [55] show that tensions arise when trying to simultaneously achieve many of these notions. However, Madaio et al. [138] emphasize that 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 [28], and persists even with perfect information about individuals. This theory is rather simplistic for most sociotechnical systems as it is essentially based on the discriminatory principle that decision-making agents derive higher utility from certain social groups.

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* [14, 157] 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 [56]. Finally, another belief-based discrimination theory is *mis-specification* [41]. Without being aware of their own bias [159], some decision makers may hold misspecified models of group differences which, in the absence of perfect information, lead to false judgment of an individual's abilities.

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 [158] 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."* [158] 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* [179] (every agent receives at least $\frac{1}{n}$ of her perceived value of resources); *equitability* [83] (every agent equally values their allocations); and *envy-freeness* [192] (every agent values her allocation at least as much as another's). While these notions capture fairness of allocations at an individual level, they treat all the individuals equally despite the differences in contrast to individual fairness notions from machine learning which rely on some similarity metric to ensure similar outcomes only for similarly deserving individuals. 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. [57] points out that one of the deficiencies 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 noticeable 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.[2]

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 $A \longrightarrow B$ 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 [120] 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 Hossain et al. [104], Hu and Chen [106], who show that adding group parity constraints can lead to a decrease in welfare for *every* group. Hu and Chen [106] show that for empirical risk minimization problems, stricter fairness constraints lead to a decrease in welfare (defined through the desired outcome of a binary classification task) for all groups in a population, as compared to having no fairness conditions imposed.

Recent works also propose fairness-to-welfare pathways that transform utility-based metrics into comparing probability of outcome [19, 104, 200], showing that fairness definitions do not automatically imply equitable outcomes from a mechanism design perspective, but on the contrary. Kasy and Abebe [121] 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, and standard definitions of fairness often fail to acknowledge this heterogeneity.

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 concepts and lessons from political philosophy, Binns [36] 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 the 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 the initial progress [202]. First and foremost, dynamic effects over time are crucial, and if neglected they can

[2] Voting theory deals with aggregating individual preferences under certain fairness axioms. As it intersects less with the ML-MD setups in our paper, we exclude discussions on voting while we acknowledge the existence of substantial works on fair voting.

worsen rather than improve the inequality and discrimination we already observe in large-scale decision-making systems. Indeed, even in simple two-stage models, Liu et al. [136] and Kannan et al. [119] highlight the possibility of harms caused by fairness constraints and the impossibility of equality; interestingly, many of those papers [119, 137] are strongly influenced by the classic economic models such as Coate and Loury [56] and Phelps [157].

Second, the type and complexity of interventions needed to achieve long-term fairness may vary significantly. For example, Hu and Chen [105] build upon the labor market model in Levin [135] 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. [196] 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 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 useful in explaining dynamic patterns of discrimination in sociotechnical systems. Bohren et al. [41] 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' misspecified bias in sequential ratings. Monachou and Ashlagi [147] build upon this theory to study the long-term effects of social bias on worker welfare inequality in online labor markets, while Heidari et al. [102] 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 [194], 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* [42, 43, 53, 61, 66, 99, 107] 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. [107] contextualize strategic investment in test preparation to falsely boost scores that are used as a proxy to quantify college readiness, and the disparate equilibria that could potentially emerge in the presence of social groups with disproportionate manipulation capabilities. Additionally, Milli et al. [144] utilize credit scoring and lending to show there is a trade-off between the utility of a decision maker who tries to protect themselves from the agents strategically modifying their features and the social burden different groups 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 [10, 124] on evaluation mechanisms and the machine learning literature [96, 143, 184] 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, with those dissimilarities having 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 [139, 156]. In this context, the uncertain relation between features and individual qualifications makes *strategyproofness* a necessity in order to make prediction-based decision making systems transparent and truly 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 [60] and disincentivizing all insurers from offering attractive insurance plans to people with specific medical conditions [204]. 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 [49, 175] and search engines [23, 76] 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 implications that their predictions cause to human subjects.

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 impacts [55, 100].

Uncertainty can also manifest with respect to outcomes, not just preferences. Many application domains consider inherently probabilistic models—for example, models of the labor market from mechanism design often consider two-stage processes in which noisy signals provide information about whether a worker is qualified or not [56, 105]. Importantly, while these models do incorporate uncertainty, the mechanism 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 impacts 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 [44], feedback loops [75], preference elicitation [39, 85, 89, 205], and explore-exploit trade-offs [38, 110, 162], 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. [8] 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 [70, 100] or representations [79, 201] to understanding causal relationships between features [122, 128]. While satisfying multiple definitions may not always be possible [125], 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 [56, 105], recent works in mechanism design [57] 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 [46] and advertising [123], 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

Most early studies of fairness in both mechanism design and machine learning propose various mathematical formulations of fairness, and normatively prescribe how fair decisions should be made. An innate but false assumption in these studies is that there is societal consensus about what fairness is. Given the impossibility to simultaneously satisfy multiple fairness notions [55, 125], decision-making systems need to be restricted to only selected principles of fairness, and selecting the right ones is a challenge especially in some critical machine learning applications in criminal justice, finance, self-driving, etc. In addition, it is also essential for the chosen principles to be socially acceptable as the resulting decision making systems can cause significant societal harm, potentially affecting everyone. 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 designing decision-making systems in a democratic way.

Machine learning research has taken steps towards this democratization goal through participatory socio-technical approaches [27, 191] to fairness. A line of work [91, 93, 131, 169, 178, 198] studies how people perceive fairness in machine learning and how to infer social acceptance on what is fair. Similar studies on fairness in mechanism design are much needed; except [132, 133], fairness perceptions have not been widely studied in mechanism design.

There are a few important questions which need to be answered before proceeding towards the democratization goal for fair automated decision making. First, *whose perceptions or assessments should be captured in the process?* While Awad et al. [17] and Noothigattu et al. [149] use crowdsourced preferences from lay humans in the famous moral machine experiment, Jaques [112], Yaghini et al. [199] have argued that preferences should be taken only from relevant individuals (e.g., primary stakeholders, ethicists, and domain experts) citing context-dependent aspect of fairness and the possible vulnerability of lay humans to societal biases which could hijack the whole process. Second, *what sort of options and information should be made available to the participants?* Some studies [17, 101, 149, 173] directly asked participants to choose the model with the best fairness notion or the best outcomes, whereas others [93, 178, 199] 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). On the other hand, Grgić-Hlača et al. [94] and Van Berkel et al. [190] study the validity of using certain input features in the decision-making process in order to achieve procedural fairness. In addition, recent literature argues that model explainability [37, 65, 161], system transparency [161, 195], and the availability of structured discussions [134, 190] can significantly help participants in making informed judgements on fairness concerns. Finally, *how should the participants' individual preferences be aggregated?* Even though most of the literature has followed some variant of majority rule for this, Noothigattu et al. [149] and Kahng et al. [116] have argued for tools like score-based bloc voting or Borda count from voting theory [73] for better representation of participants' choices.

Thus, even though the machine learning community has worked towards meeting societal expectations on fair decision-making systems, there are still a number of gaps which could be addressed with tools from voting theory. Not only can mechanism design take inspiration from such studies in machine learning, but also both fields can work together towards a better understanding of societal fairness perceptions and democratization of fairness in automated decision-making systems.

4 APPLICATION DOMAINS

In this section we discuss several application domains of machine learning and mechanism design to demonstrate the lessons of Section 3, point out gaps between the two fields as well as possibilities for bridging those gaps. The application domains exhibit a variety of relationships between machine learning and mechanism design components in practice, and underscore their complex interplay.

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. And more fundamentally, one might argue that the overall social system (e.g. the criminal justice system) in which an application (e.g. recidivism prediction) is embedded is unjust, and that this cannot be remedied by any technical fairness intervention. We discuss applications merely as illustrative of the lessons we have articulated, and 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 which 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 (usually some variant of the second-price auction [153]) 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.

Several recent studies show that the resulting ad deliveries could lead to unfair distribution of audience, i.e., users who differ on sensitive attributes such as gender [130], age [13], or race [11], can end up receiving very different types of ads. For example, searches done with black-sounding names are highly likely to be shown ads suggestive of arrest records [183]. Another study showed that women were shown relatively fewer advertisements for high paying jobs than men with similar profiles [62]. When ads are about housing, credit or employment, such disparities can eventually lead to disparate life opportunities.

One of the first reasons behind such unjustifiably skewed ad delivery can be the explicit targeting of users based on sensitive attributes [11, 77] which can be stopped by examining and 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 the availability of other personally identifiable information such as area code [176], or using a biased selection of the source audience in the Lookalike audience tool by Facebook. Following a lawsuit [177], Facebook removed targeting options for housing, credit, and employment ads [67].

Other studies [9, 171] 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, e.g., relative competition between general opportunity ads and category-specific ads for items like female fashion can result in opportunity ads being shown to more male audiences. This issue can be resolved 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) [87], or using different bidding strategies for different demographics groups [148]. However such type of targeting/bidding has been disallowed by the platforms because of earlier exploitation by discriminatory advertisers. Moreover, utility-optimizing strategic advertisers may not accept such advertiser-side solutions as they incur some cost on the advertisers. Secondly, on the auctioneer's side, the allocation mechanism can be redesigned to ensure fair audience distribution [47, 51, 69, 108]. Along with the welfare optimization goal (same as revenue optimization), group fairness constraints can be used to ensure fair audience distribution [47], and individual fairness constraint [51] or envy-freeness constraint [108] can be adopted to ensure similar individual satisfaction of the users. Most of the above studies have highly focused on the mechanism design components (advertisers' targeting/bidding strategy and the auction mechanism) of online ad delivery. However, all the three components—advertisers' strategies, platform's relevance prediction, ad allocation mechanism—could very well 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 relevance predictions. Thus, to build a fair online ad ecosystem, there is a need to study the role of relevance prediction models, and how it should be paired with suitable mechanism design components. In this regard, a line of work in machine learning based preference elicitation in auction settings [129, 154, 205]—combines both learning and design aspects—can be explored and potentially extended to online ad settings.

4.2 Admissions in education

Schools and universities increasingly rely on ML-based algorithms to inform admissions decisions [140]. Mechanism design has traditionally studied problems such as school choice, college admissions and affirmative action (e.g., [5, 6, 48, 50, 84, 86, 109, 118]). 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 (for example, all the public schools in Boston or New York may commit to a common matching process), while in college admissions, the decision process is not coordinated and 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 [78]. 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 [103, 155, 166, 168] 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 [74, 74, 96, 107, 109, 137] usually poses the admissions problem as a classification task to predict whether an applicant is "qualified" or "unqualified" to attend their university[3] 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. [74], 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. [119] 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 [125], 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 [34, 160, 197]. 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 [40, 163, 170]. While some contend that quantitative tools might help to reduce discrimination [59], others warn that hiring discrimination will not be solved by machine learning alone [90]. 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 [52]: 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 [113, 170]. Moreover, many predictions about candidates are ultimately used in contexts where there is a limited hiring capacity. As a result, predictions

[3] We do not endorse the normative ideal that someone is inherently qualified or unqualified to receive an education. However, for that reason, the example still illustrates the inequality exacerbated by algorithmic solutions to human problems.

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 [163], 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. [167] and Monachou and Ashlagi [147] describe how mechanisms that use customer ratings to evaluate workers can internalize customers' discriminatory tastes. Barzilay and Ben-David [24] call attention to the ways in which platform design can be used to create or reduce wage disparities. Similarly, Hannák et al. [97] 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. [72] 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 [26]. 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 [186] and recidivism prediction [12]. 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 [63, 82].

Since the deployment of predictive models in the criminal justice system is a contested idea [98], 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 [58, 115] while some of them are impossible to simultaneously satisfy [55, 125]. 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 [91, 93, 131, 169, 178, 198] 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 [189] 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 machine learning system regarding innocence and guilt ignores human incentives in the criminal justice pipeline, as well as 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. However, 95% of felony convictions in the United States are obtained through guilty pleas [1], and 18% of known exonerees pleaded guilty or did not contest to crimes they did not commit. Machine learning and statistical 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 [32]. 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 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 [127] and the particular context in which risk assessment tools are utilized [180].

4.5 Health insurance markets

Interactions between machine learning and mechanism design are salient for fairness in health care. For example, prior work studied how machine learning formulations may under-predict black patients' health care needs [150]. Here, we draw attention to problems that arise at the intersection of machine learning and mechanism design in health insurance markets.

The Patient Protection and Affordable Care Act (ACA) [2] was designed in part to defuse health-insurer incentives to refuse or avoid coverage to individuals with higher healthcare costs (i.e. selection incentives) [60]. One way 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 [4, 204]. 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 [88, 111].

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. At the same time, 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 response to the systematic error observed in the HCC is 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 [60].

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 [204].

We see each discipline's techniques applied separately and in a component-wise fashion towards the goal of building 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 [172] 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 [3] 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. Kallus and Zhou [117] observe that algorithms might still yield "bias in, bias out" phenomena, even with fairness constraints, resulting from the historical imbalance of loan acceptance [136, 164].

As transparency increasingly becomes a legal obligation of financial institutions, such technology is particularly susceptible to disparities [22], largely because of two tensions shaped by the incentives of different stakeholders. First, individuals who gain insight about the internal decision policies of the institutions, might have disproportionate recourse abilities, based on their current financial status and limited access to opportunity. As a result, since financial institutions are typically for-profit organizations aiming to maximize their utility, Milli et al. [144] note that a (perhaps partial) 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 being transparent towards their customers and not fully revealing their decision policies for intellectual property reasons [21, 145]. In this context, Tsirtsis and Gomez-Rodriguez [188] discuss an attempt to maximize utility, which may provide limited recourse options to disadvantaged populations in favor of majority groups. 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 [172] 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 [45, 64, 95, 141, 151]. Inequality at the level of individual connections is often reinforced by algorithms that use these connections for learning: in opinion diffusion [9, 81, 182], recommendation [181], clustering [54, 126], and others. Often, such inequality arises from the individual preference for establishing new connections as well as from pre-defined communities [15]. Recent papers discuss these patterns through the lens of welfare economics or equilibrium strategies, with Avin et al. [16] analyzing the utility function for which preferential attachment is the unique equilibrium solution in a social network. Understanding the incentives behind network creation patterns is thus crucial for designing better algorithms that learn from relational data and tackling bias at its root cause, as Section 3.2 teaches us through an understanding of long-term effects of fairness.

Beyond this, several works argue that ranking and retrieval algorithms not only reinforce existing bias, but also cause changes in people's behavior [152]. 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 payoff. 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. [18], who focus on finding stable equilibria for which users get the best pay-off for their desired items. Ben-Porat and Tennenholtz [30] 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. [25], Ben Basat et al. [29] 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. [146] 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 [49] 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, Jon Kleinberg, Hannah Li, Irene Lo, Francisco Marmolejo-Cossio, and Angela Zhou for helpful comments and suggestions. This project has been part of the MD4SG working group on Bias, Discrimination, and Fairness.

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

REFERENCES

[1] Guilty Plea Problem. *The Innocence Project.* URL www.guiltypleaproblem.org.

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

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

[4] Risk adjustment. https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/Risk 2018.

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

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

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

[8] 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.

[9] 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.

[10] 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.

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

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

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

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

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

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

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

[25] 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.

[26] 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.

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

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

[29] 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.

[30] 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.

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

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

[33] 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*.

[34] 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.

[35] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 405–414, 2018.

[36] 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.

[37] 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.

[38] 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.

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

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

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

[42] 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.

[43] 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.

[44] 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.

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

[46] 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.

[47] 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*, 2019.

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

[49] 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.

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

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

[52] 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.

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

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

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

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

[57] 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.

[58] 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.

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

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

[61] 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.

[62] 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.

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

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

[65] 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.

[66] 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.

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

[68] 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.

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

[70] 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.

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

[72] 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.

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

[74] 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.

[75] 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.

[76] 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.

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

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] Duncan Karl Foley. Resource allocation and the public sector. 1967.

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

[85] 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.

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

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

[88] 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.

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

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

[91] 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.

[92] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *Advances in Neural Information Processing Systems*, 2016.

[93] 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.

[94] 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.

[95] 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.

[96] 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.

[97] 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.

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

[99] 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, 2016.

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

[101] 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.

[102] 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.

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

[104] 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.

[105] 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.

[106] 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.

[107] 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.

[108] 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.

[109] 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, 2019.

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

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

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

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

[114] 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.

[115] 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.

[116] 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.

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

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

[119] 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.

[120] 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.

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

[122] 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.

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

[124] 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.

[125] 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.

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

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

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

[129] 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.

[130] 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.

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

[132] Min Kyung Lee and Su Baykal. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1035–1048, 2017.

[133] Min Kyung Lee, Ji Tae Kim, and Leah Lizarondo. A human-centered approach to algorithmic services: Considerations for fair and motivating smart community service management that allocates donations to non-profit organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3365–3376, 2017.

[134] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–35, 2019.

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

[136] 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.

[137] 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.

[138] 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.

[139] 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.

[140] 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.

[141] 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.

[142] 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.

[143] 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.

[144] 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, 2019.

[145] 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, 2019.

[146] 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.

[147] 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.

[148] 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.

[149] 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.

[150] 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.

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

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

[153] 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.

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

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

[156] 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.

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

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

[159] 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.

[160] 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.

[161] 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.

[162] 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.

[163] 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.

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

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

[166] 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.

[167] 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.

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

[169] 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.

[170] 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.

[171] 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.

[172] Lauren Saunders. FinTech and Consumer Protection: A Snapshot. 2019.

[173] 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.

[174] 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.

[175] 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.

[176] 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.

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

[178] 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.

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

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

[181] 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.

[182] 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.

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

[184] Behzad Tabibian, Stratis Tsirtsis, Moein Khajehnejad, Adish Singla, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. Optimal decision making under strategic behavior. *arXiv preprint arXiv:1905.09239*, 2019.

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

[186] The AI Now Institute. AI Now 2019 report. Technical report, 2019.

[187] 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.

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

[189] 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.

[190] 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.

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

[192] Hal R Varian. Equity, envy, and efficiency. 1973.

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

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

[195] 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.

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

[197] 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.

[198] 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.

[199] 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.

[200] 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, 2017.

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

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

[203] 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.

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

[205] 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.

A APPENDIX

A.1 Common definitions of fairness in machine learning

In this table, we have compiled a list of common definitions of fairness in machine learning as collected by Mehrabi et al. [142].

Name	**G/I/Other**	**Setting**	**Informal definition**
Demographic Parity [70]	Group	Binary classification	Proportion of each group achieves positive classification at equal rates.
Equality of Opportunity [100]	Group	Binary classification	*Qualified* proportion of each group achieves positive classification at equal rates.
Equalized Odds [100]	Group	Binary classification	Probability of positive classification is equal across both groups if true label is either 0 or 1.
Treatment Equality [33]	Group	Binary classification	Ratio of false positives and false negatives is equal across groups.
Fairness through awareness [70]	Individual	General	Individuals who are similar according to a given distance metric are treated similarly.
Fairness through unawareness [92]	Other	General	Algorithm does not explicitly use protected class in the decision-making process.
Counterfactual fairness [128]	Other	General	Would the individual been treated the same if they were a member of the other group.