Strategic Evaluation: Subjects, Evaluators, and Society

Benjamin Laufer
bdl56@cornell.edu
Cornell Tech
New York, New York, USA

Jon Kleinberg
kleinberg@cornell.edu
Cornell University
Ithaca, New York, USA

Karen Levy
karen.levy@cornell.edu
Cornell University
Ithaca, New York, USA

Helen Nissenbaum
hn288@cornell.edu
Cornell Tech
New York, New York, USA

ABSTRACT

A broad current application of algorithms is in formal and quantitative measures of murky concepts – like merit – to make decisions. When people strategically respond to these sorts of evaluations in order to gain favorable decision outcomes, their behavior can be subjected to moral judgments. They may be described as 'gaming the system' or 'cheating,' or (in other cases) investing 'honest effort' or 'improving.' Machine learning literature on strategic behavior has tried to describe these dynamics by emphasizing the efforts expended by decision subjects hoping to obtain a more favorable assessment — some works offer ways to preempt or prevent such manipulations, some differentiate 'gaming' from 'improvement' behavior, while others aim to measure the effort burden or disparate effects of classification systems.

We begin from a different starting point: that the design of an evaluation *itself* can be understood as furthering goals held by the evaluator which may be misaligned with broader societal goals. To develop the idea that evaluation represents a strategic interaction in which both the evaluator and the subject of their evaluation are operating out of self-interest, we put forward a model that represents the process of evaluation using three interacting agents: a decision subject, an evaluator, and *society*, representing a bundle of values and oversight mechanisms. We highlight our model's applicability to a number of social systems where one or two players strategically undermine the others' interests to advance their own. Treating evaluators as themselves strategic allows us to re-cast the scrutiny directed at decision subjects, towards the incentives that underpin institutional designs of evaluations. In practice, the moral standing of strategic behaviors often depend on the moral standing of the evaluations and incentives that provoke such behaviors. We apply our framework to a variety of extended examples and discuss ethical implications.

KEYWORDS

Strategic behavior, Evaluation, Measurement

EAAMO '23, October 30-November 1, 2023, Boston, MA, USA

ACM ISBN 979-8-4007-0381-2/23/11.
https://doi.org/10.1145/3617694.3623237

ACM Reference Format:
Benjamin Laufer, Jon Kleinberg, Karen Levy, and Helen Nissenbaum. 2023. Strategic Evaluation: Subjects, Evaluators, and Society. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23), October 30-November 1, 2023, Boston, MA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3617694.3623237

1 INTRODUCTION

An important recent theme in machine learning and mechanism design research has been its expansion into domains such as hiring, lending, educational admissions, and other settings where people apply for opportunities (a job, a loan, a place at a selective college) and an institution draws on algorithmic assistance for evaluating them. Because of the increasing interest in such applications, a growing body of work focuses on the strategic behaviors of people subjected to these types of algorithmic decisions—that is, attempts to modify one's attributes to obtain a more favorable classification in an algorithmic evaluation [5]. Recent work on this issue has explored a range of different bases for such strategic behavior, ranging from strategies that improve the attributes of interest in the evaluation (such as when a student studies the material on a test so as to score higher) to strategies that "game" the process by improving measurable features without necessarily affecting the attributes the evaluation is designed to measure (such as when a student devotes time to perfecting test-taking strategies, rather than mastering the underlying material).

The work on these issues in the computer science and mechanism design communities has thus conceived of the process as a two-player game between an evaluator and a decision subject, with the institution constructing an evaluation designed to robustly measure certain attributes of the decision subject, and the decision subject investing effort to score well on the evaluation. This two-player structure, in which the first party engages in self-presentation while the second party focuses on eliciting underlying information about them, is indeed so central to the work in this area that it is often not explicitly called out as an assumption of the models being used. But if we think about the issues that arise in the process of evaluation, it is clear that some of them are not well-explained by this type of two-player interaction. For example, sociological and organizational research has demonstrated that evaluating candidates according to their "fit" with company culture can encode underlying cultural biases and cement inequities in an organization. If a job applicant spends energy presenting themselves in a way that conveys fit in this sense, should we think of them as "gaming" the hiring process?

Or, alternatively, should we think of the hiring process itself as deficient in some way? A model that treats the evaluation merely as an elicitation device for an applicant's attributes will struggle to identify the deeper normative concerns at play in such an example.

Consider a second example: when a university assigns grades to its students, we typically describe this grading process as a way of measuring student performance. But in this narrow view, we may miss some of the other interests at play in a grading scenario, together with their normative implications. A university that is engaging in grade inflation, for instance, might find that its instructors receive higher teaching evaluations and its development program receives larger alumni donations in a regime with higher grades. Students may also view themselves as benefitting from such a regime. If both parties appear to be advantaged by the decision to assign uniformly high grades, how do we pinpoint what is intuitively undesirable about grade inflation?

These and many other examples suggest that a fuller understanding of strategic evaluation requires that we include an additional player in the model. We draw on sociological theory and empirical evidence about how evaluating institutions function in society: while evaluating institutions are frequently tasked, explicitly or implicitly, with implementing broad societal goals and values, they also operate out of self-interest, which may be more or less aligned with these societal aims. From this starting point, we observe that in a richer model of strategic evaluation, it is not only the decision subjects who operate out of self-interest: the design of an evaluation should *itself* be understood as a self-interested behavior by the institution, which aims to achieve its own goals under various social, legal, and organizational constraints. The institution is in turn held to account by a third player, which we can think of as playing the role of society—in the form of laws, regulations, norms, or individual authorities tasked with oversight.

Our introductory examples suggest that many of the central considerations in the design of evaluations are better understood as clashes within this three-player structure, between the strategic actions of the individuals being evaluated, the institution performing the evaluation, and society's expectations for what the evaluation should be achieving.

The paper is organized as follows. Section 2 provides our three-player model, aiming to capture the various ways that an evaluation outcome can diverge from societal goals. Section 3 discusses three extended examples: hiring practices, grade inflation, and sports. In Section 4, we use the model to enumerate the set of possible scenarios where the interests of the three players either align or diverge. Section 5 discusses the ethical implications of our model, aiming to recast ethical scrutiny in light of evaluators' strategic aims. We discuss further related work in Section 6.

2 OVERVIEW

We suppose that society has determined some desirable property of interest, and it would like to find the people who exhibit this property. But since society lacks the ability to perform this assessment itself, it delegates the task to an *evaluator*. The evaluator constructs a test that it gives to *subjects*, and those who pass the evaluation are deemed to satisfy the property. ("Society," of course, is not a monolithic actor with a single set of goals. As our model will illustrate, we intentionally conceive of society capaciously—encompassing both situations in which a governing body is imbued with some regulatory authority to oversee the activities of evaluators, as well as situations in which societal interests are manifested as the expression of public values, but without explicit organizational oversight.)

There are several sources of 'slippage' that are possible in this setting—that is, cases in which an assessment fails to meet its mandate or achieve broader societal goals. We would like a model that is capable of considering these discrepancies in a unified manner. The first source of slippage is the gap between someone's performance on a test and their underlying ability. A subject might have slept badly the night before a math test, leading to a score that does not reflect their skills. Or, a strategic test-taker might have chosen to invest in skills that boost their score on the assessment without improving underlying properties (i.e., some form of 'gaming'). A second source of slippage is the gap between the aim of the evaluator and the design of the evaluation. No evaluation perfectly measures its intended quantity. Myriad factors constrain the evaluator and limit the signal that an evaluation can capture, either by introducing noise or bias to the measurement. This type of misalignment has received significant attention in work on the role of *proxies* in classification: practical evaluations generally need to rely on measurable properties that stand in for the property of interest, but which do not precisely coincide with it. The third source of slippage is the gap between the evaluator's aim and the true property of interest to society. Once we appreciate that the evaluator, like the subject, is also a strategic actor with their own self-interest, then we can see that the misalignment can also arise from forms of gaming or cheating by the evaluator: the evaluator might care about a property other than the one society is interested in assessing, and they might have correspondingly created a test that is better at measuring their property than it is at measuring the one of interest to society.

There are thus multiple issues that we need to keep in focus for the purposes of our analysis: a given circumstance inhabited by the subject might or might not be sufficient to pass a test; a given way of passing the test might or might not correspond to what the evaluator is trying to measure; and what the evaluator is trying to measure might or might not correspond to the underlying societal values that motivated the test in the first place. To keep track of these issues we therefore introduce a formal model that includes all of these facets. The model cannot by itself resolve the underlying ethical questions about the behavior of the subject and the evaluator in any given situation, but it can provide precision about the nature of these situations as a starting point for ethical analysis.

The Model

In order to formalize the notion that the test is imperfectly measuring an underlying property of interest, we need to represent the idea that the true state of the subject is hidden and only partially observable. Therefore, the fundamental ingredient in our model is a set S of abstract *states*, where each state serves as a possible description of the subject. In general, we view the set of states as enormous, since the states need to be able to recognize fine-grained distinctions between subjects: if two subjects differ in a way that

might be relevant to some form of evaluation, this should imply that these two subjects reside at different states. For example, if one is evaluating a student's ability to multiply numbers, then the state should be expressive enough to describe the student's aptitude at multiplication and how they came to acquire this aptitude. Similarly, if one is evaluating an athlete via a 100-meter sprint, then the state should describe the athlete's sprinting abilities, including information about their training up to this point, as well as the conditions under which the race is run. The fact that states are expressive enough to capture fine-grained differences between subjects means that for any particular subject, it will not in general be possible to learn their precise state from any limited amount of interaction with them.

Any property can be described by the set of states at which it holds, and our discussion thus far has implicitly been concerned with four properties that can in general all be different from one another:

- Society's initial property of interest—the concept with which we began—can be viewed as a set of states $I_s \subseteq S$.
- The evaluator might have motivations that are distinct from simply assessing the property I_s; thus, we assume that the evaluator is interested in identifying subjects who belong to some possibly different set of states $I_e \subseteq S$.
- Since the state set is enormous, and states might differ in hard-to-discern ways, it is generally not possible to perfectly evaluate a particular property; as a result, the set of states that correspond to passing the evaluator's test is a set $P \subseteq S$ that might be different from both I_s and I_e.
- Finally, the subject begins at some initial state $s_0 \in S$ and has a budget of effort that they can spend to move to a new state, $s_1 \in S$, with the goal of reaching a state that passes the test. Let $R \subseteq S$ be the set of all states that the subject can reach using this budget of effort; thus, it is possible for the subject to pass the test if and only if there is a state in $R \cap P$ — both reachable and among the passing states.

Now, for any collection of subsets $C_1, C_2, \ldots, C_k \subseteq S$, let's say that two states s and s' are *indistinguishable* with respect to $C_1, C_2, \ldots, C_k$ if for each C_j, the state s belongs to C_j if and only if the state s' does. Notice that indistinguishability is an equivalence relation, and so it divides the state space into equivalence classes. Since a given state can either belong or not belong to each of $C_1, C_2, \ldots, C_k$, there are 2^k possible equivalence classes, though some of them may be empty.

Using the four subsets of the state space we have defined—I_s, I_e, P, and R—the different scenarios of interest to us can be categorized by whether a given state belongs (or does not belong) to each of I_s, I_e, P, and R; that is, there is a different scenario for each of the $2^4 = 16$ equivalence classes of indistinguishable states with respect to I_s, I_e, P, and R. These categorizations are illustrated in Figure 1. As discussed above, a state's membership in a given equivalence class does not convey normative information on its own, but this decomposition into equivalence classes provides a starting point for ethical analysis by systematically mapping the scenarios that can arise according to these four underlying dimensions.

3 EXTENDED EXAMPLES

In order to make the utility of our three-player model more concrete and to demonstrate how the interests of subjects, evaluators, and society can be variably aligned or misaligned, we discuss three example cases: hiring, grade inflation, and sports.

3.1 Hiring

In hiring, a variety of evaluations may be employed to assess candidates' fitness for a position. We can think of hiring as a multi-stage process in which an initial pool of candidates is winnowed down into progressively smaller sets; evaluations are conducted at each stage in order to select the candidates who will progress through the pipeline [7]. Recently, a good deal of research has focused in particular on initial screening steps in the hiring process. These could include algorithmic tools to analyze resumes; personality quizzes, games, and analysis of video interviews to predict a candidate's likelihood of job success; or the degree to which candidates exhibit certain qualities, like resourcefulness or grit [40]. Much attention has been paid to the fairness and diversity implications of these tools, many of which are poorly validated. Beyond initial screening stages, evaluations commonly involve candidate interviews and other face-to-face activities with hiring managers or prospective colleagues, ostensibly designed to assess aptitude and ability to respond to questions or solve problems related to the work task, or to gauge a candidate's collegiality and "fit" with workplace norms and culture.

A two-player model of strategic behavior might direct attention to how candidates seek to present themselves to the hiring firm in order to signal their aptitude or qualifications for the job, given the types of evaluations to which they expect to be subjected by the hiring firm. They might do so, for instance, by tweaking their resumes—say, by describing accomplishments using terms likely to be viewed favorably by an algorithm, accumulating "fluff" credentials, or even providing false or exaggerated information about previous accomplishments. Or they might seek to signal cultural "fit" at an in-person interview via clothing choices, comportment, or chosen topics of conversation.

Viewing the design of the evaluation, itself, as a strategic activity—and concomitantly, the evaluator (the hiring firm), itself, as a self-interested actor—foregrounds new questions. In a two-player model, a firm's evaluation is implicitly treated as representative of broader societal interests. But in a three-player model, the firm's own goals in conducting the evaluation may diverge from goals that are aligned with societal welfare or normative values. Social scientific research suggests that this divergence is not uncommon. Firms are known to implement evaluations in order to hire candidates that "fit," a notoriously slippery concept, which often manifests as demographic homogeneity with hiring managers and other current employees. Cultural "fit" with a firm may thus diverge from societal ideals of merit, fair treatment, and nondiscrimination in employment. Ray's [42] theory of racialized organizations, for example, describes mechanisms through which seemingly neutral hiring and credentialing processes in firms can be racially exclusionary despite legal antidiscrimination mandates. Rivera's ethnographic research [43] on hiring at elite firms shows that applicants from

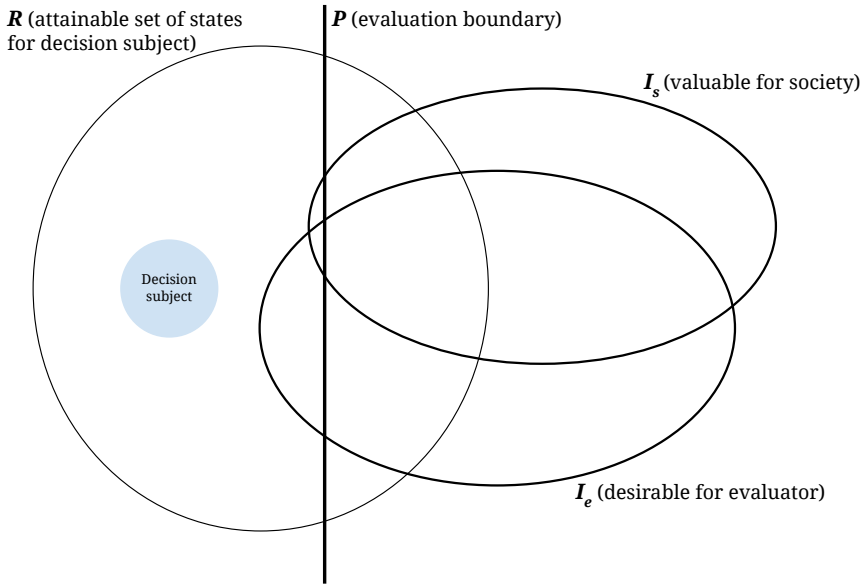

Figure 1: Diagram representing relevant states in the three-party model of evaluation. Each position for the decision subject represents a state which may or may not be described by any of the following four properties: Attainable for the decision subject (R), valuable for society (I_s), desirable for the evaluator (I_e), and passing the evaluation (P). Not all Boolean combinations are depicted.

well-resourced backgrounds do better across all stages of the hiring process due to a combination of economic advantages, social connections, and cultural resources that signal their social position to gatekeepers (i.e., hiring managers) (see also Bourdieu [9]).

Consider the following hypothetical example. In some law schools, student services staff sponsor golf instruction for law students who are unfamiliar with the game. Though golf lessons may seem orthogonal to the ostensible substantive goals of legal education, the lessons are designed to equip students with the *cultural* toolkit for job success. Golf is, traditionally, a sport strongly associated with economic privilege, and one historically off-limits to women and non-white players; as such, privileged white men are more likely to know how to play golf. Given the reality that many elite law firms are also disproportionately composed of privileged white men, and that those firms may seek to hire "the kinds of people" who know how to play golf (i.e., golf-playing is an indirect proxy for whiteness, maleness, and socioeconomic privilege), golf instruction in law schools can be understood as a means of trying to assist law students in signaling cultural fit. Indeed, law school golf programs are often explicitly directed toward women, and place emphasis on basic golf etiquette and literacy as well as skills. (For instance, participants in a golf program for women students at Arizona State University's Sandra Day O'Connor College of Law acknowledged that they were learning to play because "golf is an access issue" and that they "didn't want to be left out" of the networking opportunities that knowing how to play golf could yield [3]).

Understood through the lens of our model, we can envision a law school graduate, seeking a job at an elite firm, who is competent in the practice of law (thereby belonging to a state in I_s, society's property of interest). Imagine that our hypothetical job seeker does not come from an elite socioeconomic background, and has not played golf before. By taking golf lessons and developing a cultural facility with golf, she is able to—and does—pass the evaluator's test (that is, successfully interview for the job) in a hiring process; familiarity with golf has enabled her to reach a state that belongs to both P and R. (Even if the hiring process doesn't include an explicit golfing component, we could imagine several ways in which this cultural knowledge might surface in an interview—for example, through discussion of a recent PGA tournament, or conversation with the hiring manager about local courses.) However, if we imagine that for the elite law firm, golf skills are valued because of their traditional correlation with a particular class background, and have served as a "cultural fit" proxy for reproducing the current demographics of the firm among new hires, then our job seeker has *not* attained the evaluator's property of interest, I_e; that property is out of alignment with the others in our model.

As we've described, by considering alignments and misalignments among these states, our three-player model provides a mechanism for directing our attention to ethical implications of strategic behavior with more nuance. In a two-player model, we might simply view our job-seeker's golf lessons as strategic behavior against the evaluator's goals, and might seek ways to limit or discount its influence on the hiring process. The three-player model shows how the evaluator's interest I_e is itself out of step with both society's interest I_s and the interest of the job-seeker, who wishes to reach a state in P. Instead, golf lessons may be recast as an effort to push back against an unjust exclusionary criterion, which serves to align the subject's strategic behavior with societal objectives. As such, it may be judged as ethically acceptable.

Further, this example demonstrates an additional benefit of the three-player model. Much contemporary scholarship on fairness in hiring processes focuses exclusively on screening stages, when algorithmic tools are used to winnow down a set of candidates to a set to be "called back" for an interview. (Most social science audit studies of these processes also focus exclusively on these early hiring stages, as demonstrated by Quillian et al. [39], for both pragmatic reasons and based on research ethics considerations.) But a good deal of biased and exclusionary hiring practice—that is, misalignment of an evaluator's objective and societal objectives—is likely to occur in *interview* stages, which are often excluded from scrutiny by researchers studying the ethical dimensions of AI-driven tools or the fairness of hiring processes [35, 52]. A broader conceptualization of strategic behavior offers a more inclusive "end-to-end" view of the hiring process and the ethical dimensions thereof.

3.2 Grade Inflation

Grade inflation is a phenomenon in which the grades students are assigned for coursework tend to "inflate" (i.e., increase) over time. Empirical data demonstrate the existence of the phenomenon across colleges and universities: in one study of 200 U.S. schools [44], "A" grades comprised 43 percent of all letter grades issued in 2009, as compared to only 15 percent in 1960. Grade inflation is particularly pronounced among private colleges and universities, even controlling for student selectivity [44].

How might we understand grade inflation through the lens of strategic evaluation? We can conceive of the evaluation as the issuance of a grade (or a set of grades) to a student based on course performance. If we think of P as representing the set of states resulting in a high grade, then the student has an interest in finding a state in P and R: a reachable state in which they receive a high grade. The student presumably benefits from high grades (e.g., as a credential for a future job search). The educational institution, as evaluator, also has interests in issuing high grades: they have a reputational interest in ensuring that graduates are successful on the job market, and high grade point averages can help to set students up for such success. When grade inflation is particularly widespread, schools may find it difficult to "deflate" due to concern about harming students' career opportunities or becoming less competitive in recruiting new students [28]. Schools may also be interested in keeping students satisfied via high grades for purposes of maintaining positive alumni relations, which bear reputational and economic dividends (e.g., they might result in greater donations to the school in the future). Accordingly, the institution has incentives to make sure that the set of states I_e it approves of contain many states in P and R: reachable states conferring high grades.

Instructors, whom we can think of as acting as agents of the educational institution, have their own set of interests, which might support grade inflation. Instructors may enjoy better interpersonal relationships with students when they assign them high grades, and empirical evidence suggests that instructors who issue high grades receive better evaluations from students, which may factor into faculty's own tenure and promotion evaluations [30]. (We could also, of course, conceive of instructors and educational institutions as separate "players" in our model that are potentially *misaligned* in their interests regarding grading; we collapse them here for the sake of simplicity in illustration, since for purposes of our discussion they both have incentives for I_e to contain many states in P and R.)

Thus far, then, both the subject and the evaluator may be aligned in their interests, supporting an inflated grade regime. But society's interest—represented as I_s—may be misaligned. Indeed, we can think of the "inflation" of the grades as an enlargement of the set I_e relative to the set I_s: the institution is willing to view many more states as deserving of high grades, whereas society might want I_s to be a smaller set that has fewer states in R (corresponding to fewer states achievable by students).

There are many potential arguments for why societal interests might be poorly served by inflated grades. Grade inflation might be problematic to the extent that grades are useful tools for distinguishing among the performance of different students; if everyone gets an "A," it's not as clear which students truly excelled [29]. Similarly, some argue that grade inflation may diminish students' motivation to excel in coursework, because attaining a top grade takes relatively less work, thereby reducing students' capacity to reach their full learning potential. The ethical and social implications of grade inflation are strongly debated, both in the academic literature on the topic [14, 16, 30, 44] and within institutions facing public pressures to rein in inflation. In situations like these, interests P and abilities R of the subject are aligned with the interests of the evaluator I_e, but misaligned with societal interest I_s, as illustrated by our three-party model.

3.3 Sports

The sports world is also a useful site for illuminating these dynamics, owing to its intentionally competitive design. In sports, acceptable means of reaching a particular objective (scoring, or winning a race) are generally made explicit via a set of detailed rules promulgated by the sport's association or governing body, and subsequently enforced by referees or other officials.

Sports, therefore, are a natural setting for studying strategic behavior. A traditional analytic structure would posit that the organizer of the sporting event (the evaluator) creates a set of rules designed to measure athletes' (subjects) abilities, while athletes look for ways to gain an "edge" within the constraints of the rules. In particular, to prepare for an event, athletes train to improve speed, strength, and agility; they strategize about and prepare for likely competitive scenarios; they gather information about the strengths and weaknesses of the competition. A number of sports scandals and armchair debates involve the normative bounds of these behaviors, when such strategic efforts cross a line from gamesmanship into territory disallowed by the organizers—including doping, illegal sign-stealing, and other forms of (what is commonly perceived as) "cheating."

Consider, for example, a championship-level track and field meet. We often think of such events as having some of the most straightforward specifications—to run a certain distance as fast as possible, or to jump as far as possible—but of course they are also controlled by rules concerning allowable equipment, racing conditions, and substances (e.g. drugs) that an athlete is or isn't allowed to ingest while training or competing. We intuitively think of the "organizers" of the event as the enforcers of the rules, where the organizers represent some amalgam of the local operations of the meet and the

international governing bodies for track and field. The two-party analysis of this setting would take this collection of organizers to be the evaluator, formulating and enforcing rules that apply to the athlete as subject.

As with the other domains we've considered, this two-party interaction between the evaluator and the subject misses a number of the central issues that arise in the process of governing a sport like track and field. A salient example is the design of the track itself: at the 2021 Tokyo Olympics, a great deal of technology and money went into the creation of a "springy" track surface (i.e. rubber granules for better shock absorption) to enable the runners to increase their speed and increase their chances of breaking records [36]. This type of technology could be pushed much further than it was at the Olympic Games; what determined the limit of the track's springiness was not technology, but a sense that going too far would risk the athletes' safety, and cross a notional line that separates the act of running on a track from the act of running across a 400-meter trampoline. This notional line was therefore enforced by material-science specifications for allowable track surfaces defined by World Athletics, which governs track and field events [4].

In one sense, this example reveals a familiar kind of strategic interaction: the hosts of the track meet spend money to commission a track whose surface pushes up against the allowable specifications, and the governing body enforces rules designed to preserve the underlying intent of the activity. But in another sense, this strategic tension is happening *within* the set of parties that a simpler analysis might have grouped together as a single "evaluator." This is precisely the richer view that a multi-party analysis makes possible: athletes would like to win races, and they work strategically within the rules enforced by the event organizer and the governing body to achieve this; event organizers would like to host track meets where world records are set, and they work strategically within the rules enforced by the governing body to achieve this. The 2021 Olympics is far from the only recent high-profile example of these issues in track and field; another recent instance is the attempt (with help from Nike as part of its *Breaking2* campaign) to create approved conditions under which it would be possible for an athlete to run a marathon in under two hours [10]. It is worth noting that once we move from a two-party view to a multi-party view, there is no reason we need to stop at three parties; for example, even the governing body of track and field is motivated to create conditions in which dramatic events happen in their sport in order to attract publicity and attention. In doing so, they operate strategically: for example, choosing how to set standards within informal constraints set by further parties, including the opinion of the public and the sports media about what constitutes a reasonable format for the event. Similar considerations arise in many other sports.

A point worth highlighting is the contrast between technical restrictions on an allowable track surface and technical restrictions on allowable equipment, such as running shoes (which, like the track, are also made of rubber and designed to be springy). Though they appear similar, a key contrast is that strategic innovations in equipment are made by parties who are helping athletes; in contrast, strategic innovations in track surfaces are made by parties whom we typically think of as maintaining the integrity of the event—but whom, according to our model, we can also view as strategic actors motivated in part by their own aims.

Since there are multiple scenarios in these settings that differ in subtle ways, our formalism in terms of subsets P, R, I_e, and I_s can help clarify the distinctions among them. In applying the formalism to our examples here, we think of the underlying states as representing not only qualities about a competitor, but also different track meet scenarios and their outcomes. We focus on some definition of success — such as whether a particular world record has been broken. P is then the set of states where this success outcome occurs; R is the set of states achievable by the athlete; and I_e and I_s are the states that are acceptable to the local organizer of the event and the global governing body, respectively. In this way, we can distinguish among the interpretation of states based on whether or not they belong to each of the four sets:

- We start with the most straightforward case, in which an athlete breaks a record under conditions that are acceptable to both the event organizer and the governing body. This is simply a state in all four of P, R, I_e, and I_s.
- Now consider the following hypothetical scenario, inspired by our discussion: an event organizer commissions a highly springy track, resulting in a new world record; but the track is later found to violate the allowable material-science specification for track surfaces. This would correspond to a state in P, R, and I_e, but not I_s.
- Thus-far unsuccessful efforts to produce a marathon time under two hours using fully approved marathon conditions correspond to a different type of state: the organizers and the governing body work together to formulate a state in all three of the sets P, I_e, and I_s — i.e., an approved outcome that breaks two hours — but because human runners are not able to attain this state, it is not in the set R of states reachable by the subject. We can think of it as an open question whether — for this activity, with P corresponding to marathon times under two hours — there in fact exists a state in all four of the sets P, R, I_e, and I_s [26].
- There are other scenarios that follow almost mechanically; for example, if an athlete breaks a world record, is subsequently disqualified by the local organizers, but has their time reinstated after a successful appeal to the governing body, this corresponds to a state in P, R, and I_s, but not I_e.

4 A MECHANICAL UNDERSTANDING

In the previous section, we discussed a number of evaluation scenarios in which social dynamics lead to strategies that may or may not serve the interests of evaluators, decision subjects, and society. We now argue that our model (described in Section 2) doesn't just capture these individual scenarios, but extends to a wide range of candidate and evaluator behaviors in various domains. Where the previous parts of this section focused mainly on identifying particularly evocative examples of behaviors and strategies, we suggest here that the model can also be useful in an *enumerative* role: By varying its parameters, our model is able to portray all the possible stories we set out to describe. Given its structure as a three-party game, we can use the model to enumerate the scope of possible scenarios – both mundane and nuanced – where the interests of a subject, an evaluator, and society either align or diverge.

Would passing serve societal values? $s_1 \in I_s$	Would passing serve evaluator's interests? $s_1 \in I_e$	Passes? $s_1 \in P$	**Example states and scenarios**
Y	Y	Y	Candidate has a strong record and grew up playing golf. She passes the interview.
Y	Y	N	Candidate has a strong record and grew up playing golf. She fails because she got sick on the interview day.
Y	N	Y	Candidate has a strong record and hadn't played golf. She passes thanks to a golf program in law school.
Y	N	N	Candidate has a strong record and hadn't played golf. She fails because she had few opportunities to network.
N	Y	Y	Candidate has a weak record and grew up playing golf. She passes the interview after networking over golf.
N	Y	N	Candidate has a weak record and grew up playing golf. She fails, despite networking, due to a positive drug test.
N	N	Y	Candidate has a weak record and hadn't played golf. She passes after lying about her experience.
N	N	N	Candidate has a weak record and hadn't played golf. She fails the interview.

Table 1: A mechanistic example of different states. In this stylized story, a hiring decision is heavily dependent on an interview process in which playing golf — and an upbringing that involved time spent around golf courses — sometimes plays a significant role. We assume that this hypothetical evaluator is interested in golf-playing candidates in order to highlight the ways that an evaluator's interests can diverge from societal goals and values. The example scenarios are individual instances within a broader set of states.

Consider a particular domain where the subject of an evaluation (e.g., a job candidate or athletic competitor) is described by some state at the time she is evaluated. Recall we defined this state as $s_1 \in S$, the state a subject might occupy after exerting strategic effort. There are three qualities of this state which, we believe, capture much of the important social context for reaching descriptive or ethical conclusions about the evaluation. The first important quality, and perhaps the most straightforward, is whether the state 'passes' the evaluation (or otherwise performs favorably). This is true if $s_1 \in P$, where the group of passing states P is defined by criteria that are decided on, strategically, by the evaluator. The second important quality of state s_1 is whether it genuinely represents an example of the quality that is desired by broader social interests. Does the subject actually excel at the activity that is supposedly being tested for? If so, we would say $s_1 \in I_s$. Finally, the third important quality is whether the candidate's state s_1 serves the strategic interests of the evaluator. If passing the evaluation subject would be strategically beneficial for the evaluator, then $s_1 \in I_e$. All together, these qualities can help us characterize and make sense of the outcome of an evaluation.

For ease of exposition in this section, we leave out the (fourth) question of whether s_1 is a feasible state for the decision subject to reach (i.e. whether $s_1 \in R$). It would not be difficult to extend our discussion to include a distinction between whether s_1 belongs to R or not, but for now we focus our analysis on states in S without including this distinction.

We therefore have a taxonomy of states based on distinguishing whether a state s_1 belongs to P, whether it belongs to I_s, and whether it belongs to I_e. For each combination of these three different qualities, we provide an example scenario that intuitively satisfies the given combination. To show how all the possible scenarios can arise in a single unified setting, we situate all of them in a stylized story of a law student applying for a job at a prestigious law firm. For pedagogical purposes, we imagine that the law firm has an overtly (and in our telling, somewhat cartoonishly) biased hiring process that grows out of the discussion in Section 3.1; specifically, the firm seeks candidates who grew up in affluent circumstances, and they attempt to discern this by inviting job applicants to play a round of golf during their on-site interview visit. Our formulation is therefore deliberately extreme so that it can make clear the distinctions among different scenarios; in more nuanced situations, we would have the same formal structure but potentially a more challenging interpretive task in distinguishing among different scenarios.

Because our model contains membership in the sets P, I_s, and I_e as yes/no predicates, we do not need to be inventive to list a

range of different scenarios in which a prospective lawyer applies to such a firm; rather, we can literally build a table that mechanically enumerates all possible outcomes for these predicates. We do this in Table 1. For example, consider the hypothetical scenario where a fantastic and resourceful law student from a low-income background takes advantage of golf lessons offered through a university. If she receives a job offer in part because she was able to network with a partner over golf, her scenario represents a particular entry in our table. Since she is an otherwise fantastic lawyer, her candidacy for the job serves society's interests $s_1 \in I_s$. However, the partner's preference for networking over golf suggests a latent prejudice, in which the candidate's background does not serve the internal interests and preferences of the firm, $s_1 \notin I_e$. However, the candidate does pass, $s_1 \in P$. This scenario is depicted in the row of Table 1 corresponding $s \in P$, $s \in I_s$, $s \notin I_e$. Now imagine a similar scenario, equivalent in every way, except that the candidate does not network over golf, and instead passes the evaluation because of her strong track record and depth of legal knowledge. This more straightforward case – where the candidate does not face prejudice in the evaluative process, and the firm simply hires somebody based on their strong record – corresponds to row 1 in Table 1.

There are a few possible ways to interpret the taxonomy. Notice first that disparities between the evaluator's interests and society's (i.e. places where a state is in one of I_e or I_s but not the other) often suggest a place where the evaluator is applying a bias that is not societally beneficial. Next, we observe that absent any regulatory intervention on society's behalf, disparities between the evaluator's interests and the outcome of the test represent *noise* or *error* in the evaluation. This is because if an evaluator has full reign over the criteria and standards composing the assessment, and still a candidate faces an outcome that is out-of-step with the evaluator's interests, then this simply suggests a noisy, or error-prone evaluation. A final observation is that, generally, a good evaluation is one where all probable and feasible states s_1 pass ($s_1 \in P$) if and only if they serve societal interests $s_1 \in I_s$. In other words, a desirable evaluation is one where the values of these two predicates should match.

5 TAKING AN ETHICAL PERSPECTIVE

The framework developed so far, and expanded through the examples in the previous sections, provides several useful perspectives on the process of quantitative evaluation. First, it allows us to appreciate that evaluators can act strategically in their own interests: in other words, that gaming and strategic behavior are not only carried out by the subject of the evaluation, but by the parties designing the evaluation as well. In this way, it helps to recast normative judgments about a subject's behavior, interpreting deviations not as much from a fixed point but, instead, in light of the evaluator's own aims, which may themselves warrant scrutiny.

The framework sheds light on the disparities that may arise between the interests of the evaluator and societal interests (i.e. social welfare) more broadly, whether expressed through collections of norms or through explicit regulation. By making these disparities a central focus of our framework, we can distinguish between cases in which the evaluator makes these disparities explicit as part of the evaluation itself, and cases in which these disparities remain covert and require additional scrutiny to unearth. This distinction is crucial for work that brings AI and machine learning to bear on quantitative evaluation: current work in this space has tended to shine a spotlight on the explicit, quantitative components of decision-making processes, giving less attention to the parts of the decision-making pipeline where an evaluator's aims might be unstated, under-specified, and difficult to discern. Our model, in which the evaluator is cast as a strategic actor, opens up the possibility of turning new modeling attention to these more implicit (and possibly covert) parts of the decision process (see Barabas et al. [6]).

Our model therefore also broadens the notion of governance and regulation for an algorithmic system that makes and enforces rules. Rather than conceiving of rule-making and rule-enforcement processes as undertaken by a single monolithic entity, as much of the literature on strategic behavior has tended to do, it becomes more tractable to analyze the internal tensions that exist within this process, between multiple rule-making parties with potentially divergent interests. Attempts to address strategic behavior in algorithmic systems, via either technical or policy mechanisms, are well-served by recognizing these complexities in real-world evaluations.

Finally, our perspective has important implications for the moral scrutiny to which strategic behaviors are often subjected. A student found cheating on a math test, or an applicant embellishing a resume, might reasonably draw moral disapproval for those actions, and on its grounds warrant rejection, penalty or down-weighting. The language we use to describe such actions, e.g. "cheating," "gaming the system," or "honest effort," may convey ethical meaning, prejudging a case even when the underlying behaviors might be ambiguous. A distinctive virtue of our framing is that by conceiving of evaluators as potentially strategic actors, too, it allows the same moral scrutiny to be directed to them and not merely to those being evaluated.

Judgments about subjects depend on judgments about evaluators. Normally, the outcome or score yielded by an evaluation mechanism is taken as legitimate grounds for choosing one candidate over another, or declaring one competitor victorious over another. If an evaluation mechanism is considered sound, that is, is considered to be an effective or reliable measure of a target quality, candidates who strategically alter their features to "beat" the mechanism in ways unforeseen, or unaccounted for by the evaluator are, in the first place, presumed to be behaving unethically. Some have pointed to differences among such workarounds that justify classifying them in different ways. Miller [33], for example, classifies actions taken by a decision subject which are causally linked to the intended outcome as "improvement," otherwise, as "gaming," where the latter suggests unjustified success on a given performance metric. Greater subtlety in assessing subjects' behaviors in ethical terms seems important, but even here the focus of moral scrutiny has not shifted away from the subject of evaluation. Although the *quality* of a given mechanism may be called into question for failing effectively to measure a target, the target of a measurement itself, typically, is taken as given, a fixed point, which is assumed to align with societal values *a priori*.

Our framing treats target states as variable. In so doing it insists that the goals and methods of an evaluator are relevant factors

in assessing the moral standing of decision subjects' strategic behaviors. Examples of morally questionable measurements include discriminatory, elitist, or overly demanding hiring procedures; even including some which purport to serve values like diversity, meritocracy and fairness. Tests can play insidious gate-keeping roles, and sports standards can be mired in corruption. Acknowledging that evaluators' strategies potentially diverge from societal mandates means that it may be unjust to pin responsibility for strategic behaviors solely on a decision subject. These dynamics suggest a need for a more nuanced, contextual assessment of the behaviors of decision subjects that takes into account the legitimacy of an evaluator's aims and methods. As the vast literature on lying suggests, absolutism aside, an act of lying may range in moral standing depending on morally relevant contextual considerations [8].

We can apply these arguments to our earlier example of law schools offering golf instruction. Such a behavior is a strategic response to an existing norm among law firms—namely, that a significant amount of communication and fraternization occurs over golf. Such a norm might introduce bias into hiring and promotion processes, because people comfortable and experienced on golf courses tend to be whiter and wealthier. As such, spending time in law school learning golf, though not traditionally thought of as causally linked to successful law practice, might be justifiable given the contextual norms and evaluation criteria. Passing negative moral judgments (or, scoffing) at people learning to play golf misses the broader social forces inducing such behaviors, and can work to further exclude people not socially positioned to learn golf at a young age.

Evaluators can behave deceptively. As we noted in Section 2, there are sources of slippage between performance on a test and a true property of societal interest. These can be described succinctly by the differences between I_s, I_e and P.

When an evaluator's goal states I_e are transparent and accessible, rifts between I_e and I_s may be more easily identified. Observed discrepancies might require new forms of oversight and standards-setting so that evaluations do not skew towards states favorable to the evaluator that diverge from societal aims and values. These discrepancies show up in our example of Nike's *Breaking2* campaign, in which the company sponsored a race that aimed to enable athletes, sporting Nike sneakers, to complete a marathon in under two hours. This goal diverges from traditional marathon goals which aim towards consistency among races. Nike's highly publicized goal drew attention, by design, to the ways that it set up the race conditions to help its runners, e.g. providing pacers and positioning them to reduce wind resistance. Even though Nike's sponsored runners outpaced the standing marathon world record, the explicit highlighting of changes to these conditions helps make clear the ways in which the improved time would not have met the governing body's requirements for an official world record.

When an evaluator's goals are unclear, misleading or deceptive, however, it can be difficult to draw normative conclusions or assign responsibility. Differences between passing states P and societal goals I_s might arise because of benign and necessary practical limitations, such as measurement error, or they might arise because of blameworthy and pernicious strategic behaviors on the part of the evaluator. Undoubtedly, it is not always clear whether an evaluator is acting perniciously or in good faith. For example, academic departments, lacking diversity, could point to larger systemic issues that create barriers for under-represented minorities and claim that their interests and intentions are aligned with societal efforts to increase diversity. As non-diverse hiring persists, however, we may have cause to question whether these arguments are given in good faith.

When an institution claims to have noble goals but outcomes diverge from I_s, a possible explanation is that it is behaving strategically according to *covert* interests in I_e. By citing universal and unavoidable issues around measurement instrumentation, evaluators may be afforded some *wiggle room* to dodge normative scrutiny even when they are acting in ways that are counter to society's interests.[1] Examples include college admissions, where the underlying concept of societal interest — college aptitude — is fundamentally contested. In light of the broad disagreement over appropriate norms and standards, colleges employ concepts like 'holistic review' which are notably opaque. These strategies successfully avoid the fundamentally value-laden questions about what college aptitude is. They also afford some potentially self-dealing behavior, like earmarking applications from friends of trustees.

The main point of this section is to emphasize that strategic gaming might be tolerated or possibly even encouraged when evaluators, through their evaluation mechanisms, reward capacities that are not in alignment with societal ends and values. Furthermore, strategic gaming may be ethically justifiable to achieve a favorable outcome in competitions where the rules have not been designed in good faith. As such, it is important to remain astute to efforts by evaluators to obfuscate evaluation mechanisms that are misaligned with societal values, which reduce the capacity to identify relevant excusing conditions and also the capacity to reliably assign moral responsibility within such systems, overall.

6 FURTHER RELATED WORK

Here, we connect our work to the formidable literatures on both strategic algorithmic systems and theories of evaluation. We do not aim to provide an exhaustive review, but rather to situate our contribution and highlight relevant work. Our takeaway is that evaluation scenarios necessarily invoke *societal values*. These encoded values and commitments clarify the moral standing of strategic behaviors.

6.1 Strategic Behavior

A growing body of theoretical and applied work in machine learning focuses on dealing with strategic behavior and distribution shifts in response to algorithmic decisions. By this view, algorithms and metrics influence decisions that have an effect on the people and systems being measured, who therefore behave strategically to attain a desired outcome.

In literature on *strategic classification* [20], this dynamic is expressed using a simple game: An evaluator (player 1) first announces an evaluation scheme, and a decision subject (player 2) responds by investing some effort to alter his features and potentially change

[1] Furthering covert goals of the sort we describe might constitute manipulation, defined as *hidden influence* [50, 51].

his classification outcome. The evaluator's goal is strong performance on a metric like classification accuracy in light of potential distribution shifts caused by strategic responses, which impede the evaluator's ability to observe the decision subject's "true" underlying labels. A variety of related models have been put forward for achieving a similar goal in other statistical settings, like regression [11, 21, 48], ranking [31], and in repeated games [23, 24, 55].

The influence of mechanism design. Strategic models of ML tasks inherit a core assumption from behavioral economics and mechanism design: that social systems can be modeled as interactions among agents who behave according to rational self-interest. The tools of mechanism design have proven useful in designing systems with multiple agents to achieve certain desirable outcomes [18]. As a result, approaches drawing from the mechanism design literature tend to focus on a certain set of goals that the evaluator might have related to the integrity and effectiveness of the assessment—for example, preventing strategic behavior from occurring in the first place, or salvaging "true" signal from manipulated features.[2] These approaches typically conceive of the evaluator as solely interested in achieving a set of goals vis-a-vis the evaluated party (the decision subject).

Social costs, disparate effects. In response, a chorus of literature invokes the "social costs" involved in algorithmic evaluation. These papers point out that assessments often involve powerful institutions setting the terms for distributing welfare and directing life outcomes—for example, through credit scoring, the provision of standardized tests in education, or the use of automated assessments in hiring. In defining societal considerations, these works tend to highlight the impact of an assessment on decision subjects. Kleinberg and Raghavan [27] consider certain forms of strategic effort as utility-improving for both evaluator and decision subject. Milli et al. [34] consider decision subject effort as a social cost that should be minimized. Hu et al. [25] consider fairness in this context, finding that classification can exacerbate inequalities if decision subjects are afforded different budgets to strategically alter their features. By explicitly considering the impact of an evaluation on its subjects, these papers exemplify some of the social and ethical dimensions of evaluative settings where an evaluator's interests are misaligned with subjects'.

Writing on the regulation of algorithmic decision-making and legal requirements aimed at transparency, Cofone and Strandburg [12] find that algorithmic decisions tend to involve strategic behavior both from decision-makers and subjects. Decision-makers, who often cite undesirable 'gaming' from subjects as a reason to keep algorithms opaque, implicitly presume that that their interests align with society's. The paper makes the observation that the goals and behaviors of either player can be, plausibly, out of step with societal interests, so gaming may or may not be desirable. There is existing empirical work, too, on the contested and nuanced ethical boundaries of behaviors described by some as 'gaming the system,' especially in the context of internet platforms, where content creators use search engine optimization [54] and other practices geared towards courting viewers [37].

[2]It has been observed that applications of mechanism design can fall out of step with societal goals [22, 53]. Meanwhile, attempts to align mechanism design with broader social interests are burgeoning. See, e.g., Abebe and Goldner [1], Finocchiaro et al. [17].

In our hiring example, the practice of golf lessons arises not as malicious or manipulative efforts among decision subjects, but as a response to a practice among the evaluating law firm(s). Our model attempts to describe the fact that *both* the evaluator and the subject engage in strategic behaviors to achieve their own interests (which may, or may not, diverge from societal goals). Thus, we believe, a fundamental question that must be asked in evaluative contexts is: what are the appropriate societal goals underpinning an evaluation? Answering this question may enable society to institute mechanisms that make sure evaluations serve these goals, especially in cases when institutional interests diverge from society's. To that end, our work re-visits models of evaluation and draws conclusions not just about the appropriateness of responses, but about the appropriateness of evaluation measurements themselves.

6.2 Evaluation

Evaluations use measurements and observations to make a judgment of merit, worth or value [45, 47]. Although evaluations always involve empirical observation, not all forms of empirical observation constitute evaluation. Counting the number of yellow cars that pass on a highway is empirical measurement but is not an evaluation per se, because it does not help make a conclusion of merit, worth or significance.

Notice that many real-world settings with strategic behavior involve an evaluation. Grades measure educational aptitude. Sports measure athletic excellence. Job interviews measure skills, experience and fit. Settings involving high-stakes social decisions frequently draw from measures to assess constructs that are, at times, murky. Qualities like deservingness, or promise, or good business instincts – these can be very difficult to ascertain. Evaluations may provide institutions with a seemingly less arbitrary way of making high-stakes decisions or allocating social welfare.

When an institution is tasked with conducting an evaluation, there is typically some societal *interest*, or *value*, that a set of people wish to measure. That value can be highly contested—as with intelligence—or comparatively less contested—as with sprinting speed. The key ingredient that constitutes an evaluation and not any other sort of description is its use as a stand-in for (or operationalization of) a *value*.

A long line of scholars, especially in education and policy contexts, have developed theories and professional standards concerned with evaluating programs [2, 15, 19, 32, 46, 49]. A shared emphasis seems to be that evaluations should not unquestioningly adopt the goals of program facilitators but instead take a broader societal view. Such emphasis can be found, for example, in pushes for *stakeholder-based* approaches to evaluation [19]. A similar theme from literature on program evaluation and auditing is the need for *third-party* or external oversight in high-stakes decision-making systems [13, 38, 41]. These works make clear that internal auditing and self-evaluation often fall short in settings where firms behave in ways counter to society's interests.

We use the word evaluation to highlight that the appropriateness of strategic behaviors is tied to questions of value: Only by understanding the values underlying a particular measurement can we conclude that a (strategic) behavioral response is appropriate or inappropriate.

7 CONCLUSION

As machine learning and mechanism design expand into high-stakes social domains, they increasingly play a role in the creation of decision rules that affect people's lives. In cases where individuals respond strategically to these rules, it can be tempting to categorize these behaviors as gaming or cheating. This paper puts forward an expanded view of evaluative systems, where the decision subject isn't the only strategic actor who deserves social or ethical scrutiny. In hiring settings, for example, it is often the strategic methods and norms used to evaluate candidates that explain behavioral responses from decision subjects. In settings with grade inflation, schools and students align their interests and strategically behave in a way that undermines a broader societal measure of interest. Taking a normative perspective, we find that the moral standing of strategic behaviors often depends on the ways those behaviors are evaluated and motivated. We argue that questions about whether decision subjects are seen as 'gaming the system' need to be viewed in light of a parallel set of questions about the interests of the evaluating institution. Our expanded (three-player) model of evaluation is able to shed greater light on a variety of scenarios where certain strategies are either in line with or at odds with the interests of others.

There are a number of promising directions for future work. Though we use a three-player model to illustrate external social considerations in evaluation systems, there is no reason to stop at three. Many systems of evaluation have a recursive structure, where each evaluator might need its own third-party oversight mechanism, creating a larger vertical hierarchy of participants. In addition to this type of vertical expansion of our model, we would welcome work aimed at disentangling the bundle of values we describe as 'societal interests'—that is, a horizontal expansion in the sets of values that are juxtaposed against the actions of the evaluator. Delineating the norms and standards behind evaluations is a complex, context-dependent, and political undertaking with the potential to affect life-prospects in significant ways; failing to wrestle with this complexity may result in unfairly and illegitimately placing a thumb on the scale in favor of one or more of the stakeholders.

ACKNOWLEDGMENTS

The authors would like to thank the members of the AI, Policy and Practice group (AIPP) at Cornell University and the Digital Life Initiative (DLI) at Cornell Tech for their feedback. In particular, we thank Solon Barocas, Smitha Milli, Kenny Peng, and Malte Ziewitz for illuminating conversations, suggestions and remarks.

The work is supported in part by a grant from the John D. and Catherine T. MacArthur Foundation. Ben Laufer is additionally supported by a LinkedIn-Bowers CIS PhD Fellowship, a doctoral fellowship from DLI, and a SaTC NSF grant CNS-1704527. Jon Kleinberg is additionally supported by a Vannevar Bush Faculty Fellowship and a grant from the Simons Foundation. Helen Nissenbaum is also supported by a SaTC NSF grant CNS-1801501.

REFERENCES

[1] Rediet Abebe and Kira Goldner. 2018. Mechanism design for social good. *AI Matters* 4, 3 (2018), 27–34.
[2] Marvin C Alkin and Christina A Christie. 2004. An evaluation theory tree. *Evaluation roots: Tracing theorists' views and influences* 2, 19 (2004), 12–65.
[3] Nicole Almond Anderson. 2022. ASU Law Program helps women master the golf club as a business tool. https://attorneyatlawmagazine.com/law-school/asu/asu-law-program-helps-women-master-the-golf-club-as-a-business-tool
[4] World Athletics. 2020. Certification System Procedures. Online at www.worldathletics.org.
[5] Jane Bambauer and Tal Zarsky. 2018. The algorithm game. *Notre Dame L. Rev.* 94 (2018), 1.
[6] Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. 2020. Studying up: reorienting the study of algorithmic fairness around issues of power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona, Spain, 167–176.
[7] Miranda Bogen and Aaron Rieke. 2018. Help wanted: An examination of hiring algorithms, equity, and bias. *Upturn.org* 0 (2018).
[8] Sissela Bok. 1979. *Lying: Moral choice in public and private life*. Vintage, New York.
[9] Pierre Bourdieu. 1987. *Distinction: A social critique of the judgement of taste*. Harvard university press, Cambridge, MA, USA.
[10] Ed Caesar. 2017. The Epic Untold Story of Nike's (Almost) Perfect Marathon. *Wired* 0 (June 2017).
[11] Yiling Chen, Chara Podimata, Ariel D Procaccia, and Nisarg Shah. 2018. Strategyproof linear regression in high dimensions. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. ACM, Ithaca, NY, USA, 9–26.
[12] Ignacio N Cofone and Katherine J Strandburg. 2019. Strategic games and algorithmic secrecy. *McGill Law Journal* 64, 4 (2019), 623–663.
[13] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul, Korea, 1571–1583.
[14] Donald Larry Crumbley, Ronald E Flinn, and Kenneth J Reichelt. 2010. What is ethical about grade inflation and coursework deflation? *Journal of Academic Ethics* 8 (2010), 187–197.
[15] Peter Dahler-Larsen. 2011. *The evaluation society*. Stanford University Press, Stanford, CA, USA.
[16] Ilana Finefter-Rosenbluh and Meira Levinson. 2015. What is wrong with grade inflation (if anything)? *Philosophical Inquiry in Education* 23, 1 (2015), 3–21.
[17] Jessie Finocchiaro, Roland Maio, Faidra Monachou, Gourab K Patro, Manish Raghavan, Ana-Andreea Stoica, and Stratis Tsirtsis. 2021. Bridging machine learning and mechanism design towards algorithmic fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual, 489–503.
[18] Sanford J Grossman and Oliver D Hart. 1992. *An analysis of the principal-agent problem*. Springer, New York.
[19] Egon G Guba and Yvonna S Lincoln. 1989. *Fourth generation evaluation*. Sage, Newbury Park.
[20] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*. ACM, Cambridge, MA, USA, 111–122.
[21] Keegan Harris, Hoda Heidari, and Steven Z Wu. 2021. Stateful strategic regression. *Advances in Neural Information Processing Systems* 34 (2021), 28728–28741.
[22] Zoë Hitzig. 2020. The normative gap: mechanism design and ideal theories of justice. *Economics & Philosophy* 36, 3 (2020), 407–434.
[23] Bengt Holmström. 1999. Managerial incentive problems: A dynamic perspective. *The review of Economic studies* 66, 1 (1999), 169–182.
[24] Lily Hu and Yiling Chen. 2018. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*. ACM, Lyon, France, 1389–1398.
[25] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. 2019. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, Atlanta, GA, USA, 259–268.
[26] Michael J Joyner. 1991. Modeling: optimal marathon performance on the basis of physiological factors. *Journal of applied physiology* 70, 2 (1991), 683–687.
[27] Jon Kleinberg and Manish Raghavan. 2020. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC)* 8, 4 (2020), 1–23.
[28] Sonali Kohli. 2014. Princeton is giving up ground in its fight against grade inflation. https://qz.com/277288/princeton-is-giving-up-ground-in-its-fight-against-grade-inflation
[29] Alfie Kohn. 2019. Why Can't Everyone Get A's? https://www.nytimes.com/2019/06/15/opinion/sunday/schools-testing-ranking.html
[30] Wayne Lanning and Peggy Perkins. 1995. Grade inflation: A consideration of additional causes. *Journal of Instructional Psychology* 22, 2 (1995), 163.
[31] Lydia T Liu, Nikhil Garg, and Christian Borgs. 2022. Strategic ranking. In *International Conference on Artificial Intelligence and Statistics*. PMLR, Virtual, 2489–2518.
[32] Donna M Mertens and Amy T Wilson. 2018. *Program evaluation theory and practice*. Guilford Publications, New York.
[33] John Miller, Smitha Milli, and Moritz Hardt. 2020. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*. PMLR, Virtual, 6917–6926.

[34] Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. 2019. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, Atlanta, GA, USA, 230–239.
[35] Arvind Narayanan. 2022. The limits of the quantitative approach to discrimination.
[36] Tariq Panja. 2021. A Track Built for Speed Is Already Producing Records. *New York Times* 0 (2 August 2021).
[37] Caitlin Petre, Brooke Erin Duffy, and Emily Hund. 2019. "Gaming the system": Platform paternalism and the politics of algorithmic visibility. *Social Media+ Society* 5, 4 (2019), 2056305119879995.
[38] Michael Power. 1997. *The audit society: Rituals of verification*. OUP Oxford, Oxford.
[39] Lincoln Quillian, John J Lee, and Mariana Oliver. 2020. Evidence from field experiments in hiring shows substantial additional racial discrimination after the callback. *Social Forces* 99, 2 (2020), 732–759.
[40] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. ACM, Barcelona, Spain, 469–481.
[41] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. 2022. Outsider oversight: Designing a third party audit ecosystem for ai governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. AAAI/ACM, Oxford, United Kingdom, 557–571.
[42] Victor Ray. 2019. A theory of racialized organizations. *American Sociological Review* 84, 1 (2019), 26–53.
[43] Lauren A Rivera. 2016. Pedigree. In *Pedigree*. Princeton University Press, Princeton, NJ, USA.
[44] Stuart Rojstaczer and Christopher Healy. 2012. Where A is ordinary: The evolution of American college and university grading, 1940-2009. *Teachers College Record* 114, 7 (2012), 1–23.
[45] Michael Scriven. 1991. *Evaluation thesaurus*. Sage, Newbury Park.
[46] Michael Scriven. 1999. The fine line between evaluation and explanation. *Research on social work practice* 9, 4 (1999), 521–524.
[47] Michael Scriven. 2007. The logic of evaluation. *Dissensus and the Search for Common Ground* 1 (2007), 1–16.
[48] Yonadav Shavit, Benjamin Edelman, and Brian Axelrod. 2020. Causal strategic linear regression. In *International Conference on Machine Learning*. PMLR, Virtual, 8676–8686.
[49] Daniel L Stufflebeam, Anthony J Shinkfield, Daniel L Stufflebeam, and Anthony J Shinkfield. 1985. An analysis of alternative approaches to evaluation. *Systematic Evaluation: A Self-Instructional Guide to Theory and Practice* 1 (1985), 45–68.
[50] Daniel Susser, Beate Roessler, and Helen Nissenbaum. 2019. Online manipulation: Hidden influences in a digital world. *Geo. L. Tech. Rev.* 4 (2019), 1–45.
[51] Daniel Susser, Beate Roessler, and Helen Nissenbaum. 2019. Technology, autonomy, and manipulation. *Internet Policy Review* 8, 2 (2019), 1–22.
[52] Briana Vecchione, Karen Levy, and Solon Barocas. 2021. Algorithmic auditing and social justice: Lessons from the history of audit studies. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, Virtual, 1–9.
[53] Salomé Viljoen, Jake Goldenfein, and Lee McGuigan. 2021. Design choices: Mechanism design and platform capitalism. *Big data & society* 8, 2 (2021), 20539517211034312.
[54] Malte Ziewitz. 2019. Rethinking gaming: The ethical work of optimization in web search engines. *Social studies of science* 49, 5 (2019), 707–731.
[55] Tijana Zrnic, Eric Mazumdar, Shankar Sastry, and Michael Jordan. 2021. Who Leads and Who Follows in Strategic Classification? *Advances in Neural Information Processing Systems* 34 (2021), 15257–15269.