

Fair Machine Learning Under Partial Compliance

Jessica Dai¹, Sina Fazelpour², and Zachary C. Lipton³

¹Brown University

^{2,3}Carnegie Mellon University

jessica.dai@brown.edu, sinaf@andrew.cmu.edu, zlipton@cmu.edu

November 10, 2020

Abstract

Typically, fair machine learning research focuses on a single decisionmaker and assumes that the underlying population is stationary. However, many of the critical domains motivating this work are characterized by competitive marketplaces with many decisionmakers. Realistically, we might expect only a subset of them to adopt any non-compulsory fairness-conscious policy, a situation that political philosophers call *partial compliance*. This possibility raises important questions: how does the strategic behavior of decision subjects in partial compliance settings affect the allocation outcomes? If $k\%$ of employers were to voluntarily adopt a fairness-promoting intervention, should we expect $k\%$ progress (in aggregate) towards the benefits of universal adoption, or will the dynamics of partial compliance wash out the hoped-for benefits? How might adopting a global (versus local) perspective impact the conclusions of an auditor? In this paper, we propose a simple model of an employment market, leveraging simulation as a tool to explore the impact of both interaction effects and incentive effects on outcomes and auditing metrics. Our key findings are that at equilibrium: (1) partial compliance ($k\%$ of employers) can result in far less than proportional ($k\%$) progress towards the full compliance outcomes; (2) the gap is more severe when fair employers match global (vs local) statistics; (3) choices of local vs global statistics can paint dramatically different pictures of the performance vis-a-vis fairness desiderata of compliant versus non-compliant employers; and (4) partial compliance to local parity measures can induce extreme segregation.

1 Introduction

Responsible implementation of any allocation policy requires robust foresight about its likely impacts. In order to be useful, such an analysis needs to take into account existing and emerging inter-dependencies between the policy and environmental factors that shape the policy’s long-term, situated consequences [20, 27]. However, to date, most studies of the performance and bias of algorithms applied to allocation decisions examine the algorithm in isolation, ignoring the wider deployment context. As a result, these analyses risk distorting our understanding of the impacts of specific algorithms, and limit our ability to anticipate broader societal implications of algorithmic decision-making.

Recently, a more critical thread in algorithmic fairness scholarship has called for a broader, systems-level approach to “fairness”, recognizing that algorithmic decisions do not happen in a vacuum [30, 28, 37, 47, 21, 25, 32]. Decisions may have long-term ramifications for individual welfare beyond the snapshot captured at

the time of prediction [37, 17]. Thus, shifting attention towards the agency, impacts, and responsibility of decisionmakers in context is imperative.

In this paper, we adopt such a systems-level approach to explore the setting where multiple decisionmakers interact in a single labor market. Rather than considering the fairness of policies that a single decisionmaker might choose (i.e. the fairness of a single algorithm), we assume that there are several decisionmakers, whose decisions impact each another via market dynamics. While there are many possible settings in which a multi-decisionmaker scenario could take place—the provision of loans, for instance—we use the job market as a toy model for this scenario, both for simplicity and to set our work in dialogue with the broader labor economics literature addressing discrimination and partial compliance.

Two factors complicate the situation. First, employers vary in terms of their hiring policies, especially concerning their adherence to fairness-promoting measures. This situation of *partial compliance* reflects the current reality of predictive algorithms in hiring, which is characterized by heterogeneity across vendors regarding the type of measures, if any, enforced for counteracting bias [45]. Second, complicating matters further, differences in hiring policies across institutions can *incentivize* strategic applications, altering the distribution of candidates subsequently seen by employers [20].

We investigate these dynamics using simulation tools. Our models consist of two types of agents: applicants and employers. The applicants each have a single “score” reflecting their perceived skill levels, and belong to one of two demographic groups: one which has been historically disadvantaged, and associated with lower scores on average, and one which has been historically advantaged, which has higher scores on average. In this work, we take no position on the extent to which this disparity is the result of systematic biases in the appraisal of the disadvantaged group, or is an accurate reflection of skills that vary because of upstream discrimination in society. Our general observations apply in both cases.

The employers may either be fairness-conscious (*compliant*)—taking into account considerations of demographic parity [12, 22], or fairness-agnostic (*non-compliant*)—deciding solely on the basis of scores. We also explore settings where applicants decide the type of institutions to which they apply strategically in light of the different incentives afforded by these selection policies. Our operationalization of fairness in terms of demographic parity is not intended as an endorsement of this measure as the appropriate fairness measure in hiring settings. Rather, our choice is based on the widespread use of the measure in current practice [45], perhaps due to a perceived connection between the quantitative measure and disparate impact doctrine in the United States and indirect discrimination regulations in the European Union [2]¹.

We emphasize that our model is not intended as a realistic depiction of the labor market. We do not claim to offer direct policy prescriptions. Instead, our purpose is to propose the simplest conceivable model that captures the effects of partial compliance. By elucidating some basic qualitative insights regarding the impacts of partial compliance in allocative decisions, we aim to clarify the associated set of concerns that must be accounted for by any regulator. We argue that if even the most simple models evidence the complex interactive effects introduced by partial compliance behavior, then these effects must be considered when discussing the impact of specific policies or algorithmic approaches.

In particular, our findings are as follows. Even with the simplest of assumptions, the relationships between the number of compliant institutions and various relevant metrics exhibit interesting phenomena:

¹See Lipton and Steinhardt [36] and Wachter et al. [54] for critical perspectives on the connection.

1. Partial compliance (by $k\%$ of employers) can result in far less than proportional ($k\%$) progress towards the full compliance outcomes.
2. The gap is more severe when compliant employers enforce demographic parity to match global (vs local) statistics.
3. Choices of local vs global statistics can paint dramatically different pictures of the performance, vis-a-vis fairness desiderata of compliant (versus non-compliant) employers.
4. When coupled with incentive effects, partial compliance can induce extreme segregation across institutions.

Our results illuminate a critical shortcoming in current approaches to understanding fairness in algorithmic-based allocations, and have significant implications for how we think about auditing decisionmakers and assessing the potential benefits of regulation. For example, simulations with our model show that even if a large fraction of employers voluntarily comply with a fairness-promoting policy, that does not necessarily mean that a commensurate fraction of the benefit (relative to universal adoption) has been realized. Consequently, a regulator assessing the urgency of implementing fairness measure should take into account that even if only 20% of the population are non-compliant with a particular voluntary measure, they may be obstructing a much larger share, say 50% of the possible benefits of the policy. Moreover, our findings suggest that in order to understand an employer’s performance vis-a-vis fairness desiderata, it is not enough to look at statistics calculated based on the stream of candidates that apply to them—we must also consider the way that the set of applicants that they encounter may diverge from the demographics of the general population, and how these dynamics involve both interactions among the employers and strategic behavior among applicants.

The rest of this paper is organized as follows. In Section 2, we survey literature from philosophy, (labor) economics, and the fair machine learning community, making connections to other work showing that the (partial) compliance among multiple decisionmakers is an essential consideration for assessing both moral responsibility and implementing practical measures. In Section 3, we establish the details of our model, including the parameters to our simulation and several axes of variation that we explore. In Section 4, we discuss our experiments and key results from those experiments. Finally, Section 5 provides a more critical discussion, including implications for regulating machine learning in allocative settings.

2 Related Work

This work builds on several lines of research in economics, fair machine learning, political philosophy, and computational social science. An extensive literature in economics models discrimination in employment. Becker [9] introduced the notion of *taste-based discrimination*, where employers’ distaste for hiring employees from a certain group results in them behaving as though hiring a worker from the marginalized group was associated with a higher cost (a “disutility”), despite workers from both groups being identical in terms of true skill level. Becker also shows that this differential treatment among employers induces a sorting of minority employees into the least discriminatory employers, with the equilibrium wage determined by the disutility associated with the marginal discriminator. While our setup and motivation are different from Becker’s, with employers intervening to mitigate (rather than instigate) disparities, this segregation effect induced by differential treatment across employers also appears in our model.

Arrow et al. [5] famously criticized Becker’s model, suggesting that discrimination thus characterized would decrease competitiveness and be driven out of the market, suggesting instead to focus on models of

discrimination driven by imperfect information. Along these lines, Phelps [44] introduced a *statistical* model for discrimination in hiring, in which it is caused by differences in the difficulty of accurately measuring the true skill level of each group of employees. Aigner and Cain [1] build on this idea, emphasizing that economic discrimination ought to be measured by differential treatment based on true skill. By contrast, we take no position on whether observed scores accurately reflect the employee’s true skill level. Finally, Coate and Loury [16] address the long-term efficacy of affirmative-action policies, finding that, depending on specific parameter settings in their model, affirmative action can either eliminate stereotypes, or appear to confirm (untrue) negative stereotypes. As our “fairness intervened” models are functionally affirmative-action policies, we also explore the long-term dynamics of such policies. However, our work focuses on the impacts of many employers adopting different policies on binary hiring decisions, not on concerns regarding stereotypes or wages.

Another related line of work calls for more realistic assumptions about the social context of allocation [30, 37, 21, 47]. In the fair machine learning literature, Hu and Chen [30] called attention to dynamics of employer-employee interactions, modeling the labor market as a series of principal-agent interactions. Hu and Chen [30] also draw upon the same threads of the economics literature, and focuses on reputation and effort exertion. Liu et al. [37] focuses on credit ratings, showing that with a simple but reasonable set of assumed dynamics, certain fairness interventions can harm the very groups they are intended to protect. Hardt et al. [28], Milli et al. [40], Hu et al. [31], Kleinberg and Raghavan [34] all focus on the strategic behavior of individuals subject to automated decisions. While these works both recognize the problem of framing decisions as classifications, none focus on the issues of partial compliance central to this paper.

By contrast, we focus on two aspects of deployment dynamics that, though critical in shaping the ethical impact of algorithms in context, tend to be abstracted away in standard evaluations of algorithmic systems. First, our model represents potential differences among decision-makers in adherence to ethical or legal obligations, thus relaxing the assumption of a central decision-maker (or, equivalently, of full compliance), according to which all relevant agents comply with what justice demands of them. Present in many philosophical theories of justice and implicitly assumed by many works in fair machine learning [21], the full compliance assumption enables one to focus theorizing on the obligations that are the “fair share” of any agent. Nonetheless, recent philosophical works have cast doubt on whether theories developed under this assumption can provide sufficient practical guidance for agents in the actual world characterized by partial compliance [4, 53]. This line of work considers when and how in circumstances of partial compliance agents might face obligations that differ from what would have been their fair share, had others complied [53, 39, 46]. In the related labor economics literature, papers tend to focus on determining the incentive structures that promote or impede compliance with regulations such as minimum wage laws [6, 49], examining their macro-level impacts on the treatment of “non-favored” groups [15].

Second, in our models, decision subjects are represented as agents capable of responding strategically to the incentive structure of the environment. While abstracted away in most analyses of algorithmic reliability, this type of secondary effect is widespread in real-world allocation settings, and achieving foresight about its impacts is a priority for policy makers [20]. Our work contributes to similar efforts in fair machine learning literature towards broadening the scope of analysis to include these effects [38, 28, 17]. Moreover, in exploring the impact of these dynamics, our work goes beyond assessments of algorithmic performance in static settings, furthering research on the long-term impact of proposed interventions [30, 29, 37].

We also build on recent research using simulation models to study fairness in ML systems [17]. While comparatively new in fair machine learning, simulation studies represent a core methodology in economics

and sociology [10, 14], and are increasingly used by philosophers to study social dynamics in general [56] and fairness in particular [42, 41]. Simulations are favored in these domains owing to their ability to model emergent outcomes of multiple interdependent decisions in non-stationary settings. Furthermore, particularly in the presence of heterogeneity in individual characteristics, simulations can yield insights that are not readily available in traditional aggregate models, such as those based on closed form solutions and/or systems of differential equations [33].

3 Simulation Setup

We now provide a detailed description of the models explored in our simulations. In all of our models of a job market with partial compliance, all applicants have exactly two attributes: (i) a score, representing perceived skill for the job; and (ii) a group identity. Applicants may belong either to the advantaged group with higher mean score (group A) or the marginalized group with lower mean score (group B). Across our experiments, we consider two levels of representation in the broader population: one where the disadvantaged group constitutes 25% of the populations and another where they constitute 50%. Our market contains a number of employers ($n_{\text{employers}} = 50$), each of whom may either be compliant, or non-compliant, hiring strictly according to score.

At each time step, some number of new applicants ($\text{new_per_step} = 1250$) enter the job market. Each newcomer to the applicant pool is randomly assigned a group membership (according to population demographics). Each applicant’s score is drawn from a normal distribution with variance 1, and mean of 0 for group A and mean of -0.3 for group B. Then, every applicant chooses one employer to apply to, and each institution hires $\text{num_spots} = 10$ applicants. Once hired, applicants are removed from the market. Additionally, we remove applicants that have not been hired after 10 rounds.

3.1 How do institutions choose applicants?

We consider three possible policies that institutions may adopt when choosing applicants to hire: one generic non-compliant strategy, and two possible fairness-conscious (i.e. “compliant”) strategies, which satisfy some version of demographic parity.

1. *Generic strategy.* Non-compliant employers simply sort all applicants received by score, and hire the top applicants.
2. *Local parity strategy.* Compliant employers with the local parity strategy satisfy demographic parity with respect to the demographics of their applicant pool at that round; in most cases, this is not the same as the overall demographics of the environment. For example, if 15% of applicants to a local-parity employer were from group B, then 15% of the employer’s hires will be from group B, even if group B comprises 25% of the entire population.
3. *Global parity strategy.* Compliant employers with the global parity strategy satisfy demographic parity with their hires, with respect to global demographics; this may or may not be the same as the demographics of their applicant pool. For example, if 25% of the population belonged to group B, even if they accounted for 35% of applicants to a global-parity employer, they would only account for 25% of their hires.

The latter two parity strategies are probabilistic—hiring $x\%$ from group B *in expectation*—rather than deterministically hiring a specific number from group B based on a rounded proportion of available

headcount. For simplicity, we only consider scenarios in which all compliant employers adopt the same strategy (either local or global).

3.2 How do applicants choose institutions?

We also consider three possible strategies that applicants may employ when choosing institutions to apply to. Let $p_{\text{compliant_g}}$ represent the probability of an applicant from group G (either A or B) choosing to apply to a *compliant* institution, scaled by the total number of compliant institutions.

1. *Completely at random.* All applicants from both groups are equally likely to apply to institutions of either type; hence, $p_{\text{compliant_a}} = p_{\text{compliant_b}} = \frac{n_{\text{compliant}}}{n_{\text{employers}}}$. This reflects *no* strategic behavior, i.e., applicants have no sensitivity to incentives.
2. *Static preference.* Over the course of the simulation, all applicants from group A have a fixed preference for applying to a non-compliant institution, and all applicants from group B have a fixed preference for applying to a compliant institution; hence, $p_{\text{compliant_a}} < \frac{n_{\text{compliant}}}{n_{\text{employers}}} < p_{\text{compliant_b}}$. This reflects strategic behavior, where applicants have some knowledge about the nature of institution policies, but no access to additional information over the course of the simulation—that is, applicants are sensitive to incentives but have limited knowledge of the system.
3. *Dynamic preference.* Over the course of the simulation, $p_{\text{compliant_a}}$ and $p_{\text{compliant_b}}$ are adjusted for each round based on the results of the previous round. For each group, if that group’s acceptance rate in *compliant* institutions is greater than its acceptance rate in *non-compliant* institutions, then the log odds ratio $\ln\left(\frac{p_{\text{compliant}}}{1-p_{\text{compliant}}}\right)$ is increased by a constant amount $\text{stepsize} = 0.05$; otherwise, it is decreased by the same amount. Equilibrium for each group is reached when the probability of being accepted at a parity institution is the same as the probability of being accepted at a generic institution. This reflects strategic behavior where applicants have access to new information at each timestep, and are able to update their strategy accordingly.

Like the employer policies, these strategies are stochastic.

4 Simulation Results

In all of our experiments, we vary the number of compliant institutions (out of 50 total) from 0 to 50. For each number of compliant institutions, we run ten trials of the simulation. For each trial, we run the simulation until it reaches equilibrium: 100 steps for static applicant strategy, and 200 steps for adaptive applicant strategy. We then continue running the simulation for the same number of additional timesteps and calculate statistics from each trial based on the post-equilibrium timesteps. In all of our plots, one dot reflects the statistics calculated from a single trial.

Sublinear gain Our first key finding is that when employees apply strategically, then under partial compliance, the aggregate benefit from an additional compliant employer depends strongly on how many institutions are already compliant. In Figure 1, all employees apply with the strategy of static preference: that is, knowing that compliant employers are more likely to hire Group B applicants, and that non-compliant employers are more likely to hire Group A applicants, employees from Group B apply to compliant employers with probability 0.55 (scaled by number of each type of employer) and employees from group A apply

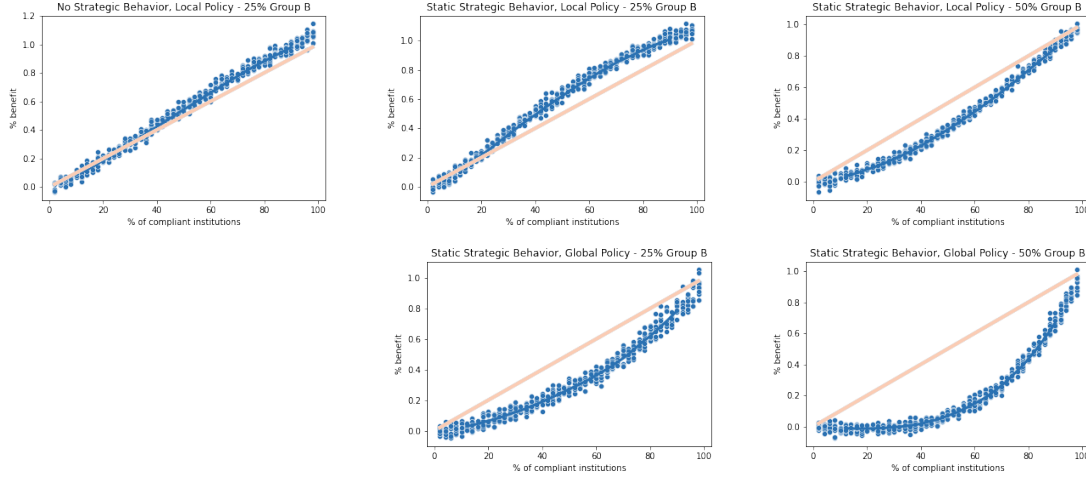


Figure 1: Benefit as measured by demographic parity. Top left plot shows market where all applicants pick employers uniformly at random. In all other plots, applicants choose a more favorable employer (compliant for group B, non-compliant for group A).

to non-compliant employers with probability 0.55. The y-axis is scaled demographic parity, where $y = 0$ corresponds to the disparate impact score $\frac{P(\text{hired}|B)}{P(\text{hired}|A)}$ when all employers are non-compliant (with our main experimental parameters, this is approximately 0.75), and $y = 1$ corresponds to “perfect” parity. One might hope that $k\%$ compliance would correspond to at least $k\%$ of the benefits, a condition that we denote *linear gain*. In Figures 1 and 2, this is illustrated by the light peach line.

Notably, when all compliant institutions satisfy fairness with respect to *global* statistics, the partial compliance curve is convex, illustrating *sublinear* gain— $k\%$ compliance always gives less than $k\%$ of the attainable benefit. Under local parity policies, the partial compliance curve can actually reflect *superlinear* gain, as when Group B constitutes 25% of the population. However, when Group B constitutes 50% of the population, these dynamics change: local parity policies now also induce *sublinear* gain, and the global parity curve indicates a more pronounced sublinear gain. In both cases, following the *global* parity policy leads to comparatively worse gains than following the *local* parity policy—that is, for any given $k\%$ compliant institutions, the percent benefit when employers satisfy global parity is lower than when employers satisfy local parity.

Static vs adaptive applicant strategy When employees are able to update their application strategy at each timestep, interesting dynamics emerge (Figure 2). Recall that the likelihood of employees from a given group applying to each type of employer (compliant vs non-compliant) is adjusted based on group-wise acceptance rates from the previous timestep. Hence, equilibrium for each group is reached when that group encounters the same acceptance rate from both compliant and non-compliant employers. Under global parity policies, the first 80% of compliant institutions are only able to push the macro-level statistics around *halfway* to parity; the remaining 50% of benefits relies entirely on the last 20% of employers to become compliant. Interestingly, under local parity policies, the first 20% of compliant employers have functionally no effect on the macro-level view of fairness, while parity is completely reached by the time around 30% of

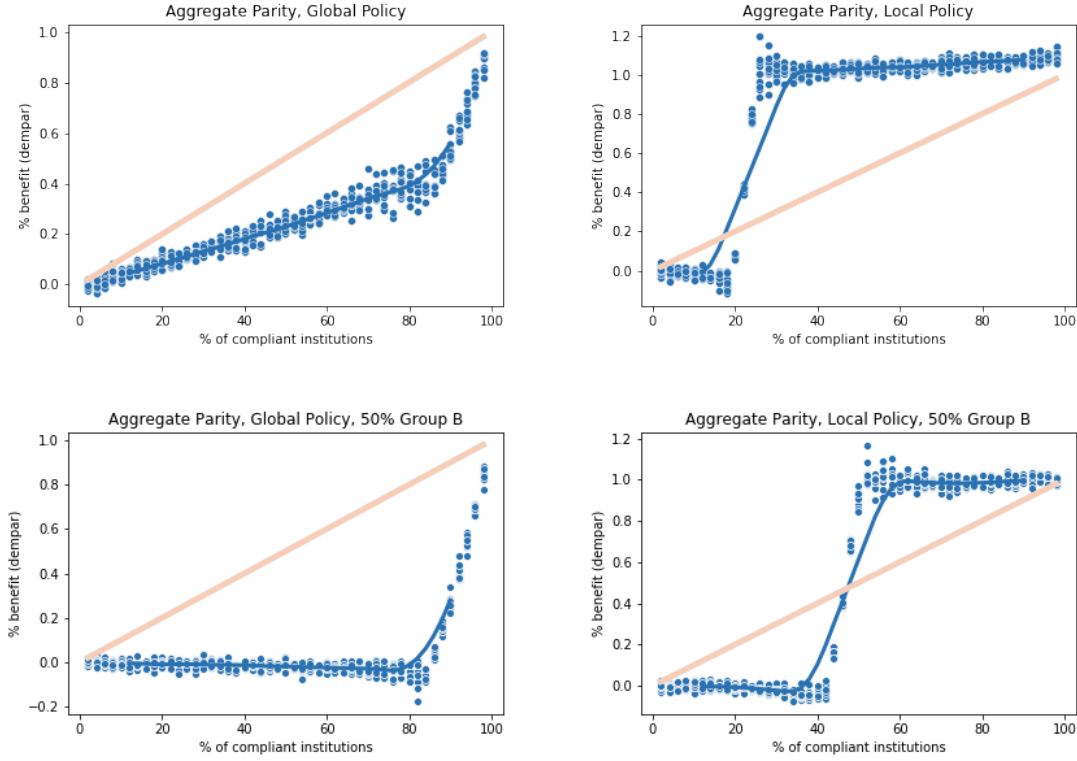


Figure 2: Aggregate statistics under *adaptive* applicant strategy, where benefit is measured by overall demographic parity. Top row: 25% Group B; bottom row: 50% Group B

employers are compliant.

For intuition as to why this is the case, we can look at the equilibrium probabilities for applying to either type of employer. Figure 3 shows that under local parity policies, the equilibrium $p_{\text{compliant}}$ (probability of applying to a compliant employer) for Group B quickly goes to 1. With 20% or more compliant employers, Group B always applies almost exclusively to compliant institutions. Meanwhile, until 26% or more employers are compliant, Group A applies almost exclusively to non-compliant institutions. Under global parity policies, the difference in preference induced by partial adoption of the fairness-promoting policy is less severe.

The emergent demographic composition of institutions A closer look at institution-specific outcomes reveals that at equilibrium, strategic applications can result in *homogeneity* within institutions and *segregation* across institutions. In the case of global parity policies, the dramatic increase in aggregate parity (Figure 2, left column) is coupled with a precipitous drop-off in the percentage of hired applicants belonging to group B in non-compliant institutions (Figure 4, bottom left). The situation is even more dire under local parity policies, as the the equilibrium strategies mean that non-compliant institutions have *no* hired

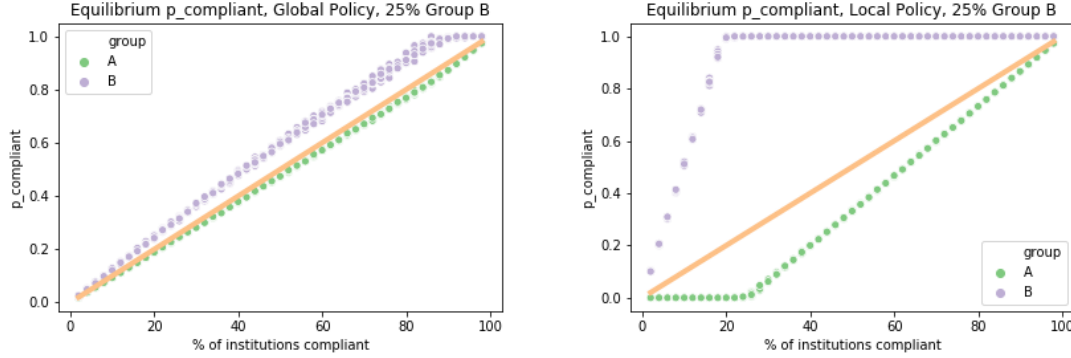


Figure 3: Groupwise *equilibrium* probability of applying to either compliant or non-compliant employers, under adaptive applicant strategy. The orange line indicates the probability reflecting no preference (i.e. probability determined solely by the number of compliant institutions currently in the system). Left: global parity policy; right: local parity policy.

applicants (or indeed, applications) from members of Group B (Figure 4, bottom right). Notably, though the aggregate parity curves under the global policy do not look so different in Figure 1, the segregation effects do *not* occur when applicants operate under a *static* application strategy: while partial compliance has some impact on the overall demographic composition of hired employees, the percentage of Group B never approaches zero (Figure 4, top row).

The impact of the original demographic makeup on adaptive applicant strategy In Figure 1, where employees were applying to firms under a *static* strategy, the impact of changing from a scenario where Group B is 25% of the population to one where Group B is 50% of the population, while significant, affects aggregate statistics in similar ways at all levels of compliance and for both global and local parity policies. However, when applicant strategies are adaptive, as in Figure 2, increasing the proportion of Group B in the population means that under global parity policies, the first 80% of compliant institutions—despite reaching 50% of the benefit when Group B was 25% of the population—actually have *no* impact on aggregate demographic parity. The critical tipping point, however, remains the same, at 80% compliance. Under local parity policies, on the other hand, the overall shape of the aggregate parity curve remains the same—two large regions with either zero or perfect parity, and one small intermediary transition region—but when Group B comprises 50% of the population, the critical transition region is between 40%-50% compliance, rather than 20%-30% compliance.

5 Discussion

Our simulations illustrate several fundamental but commonly overlooked issues that plague the ethical evaluation and governance of algorithmic tools in consequential allocation settings.

Beyond narrow assessments of fairness: diversity and integration Consider first that, in many allocative contexts, task-related utility and fairness are not the only desiderata. For example, in hiring

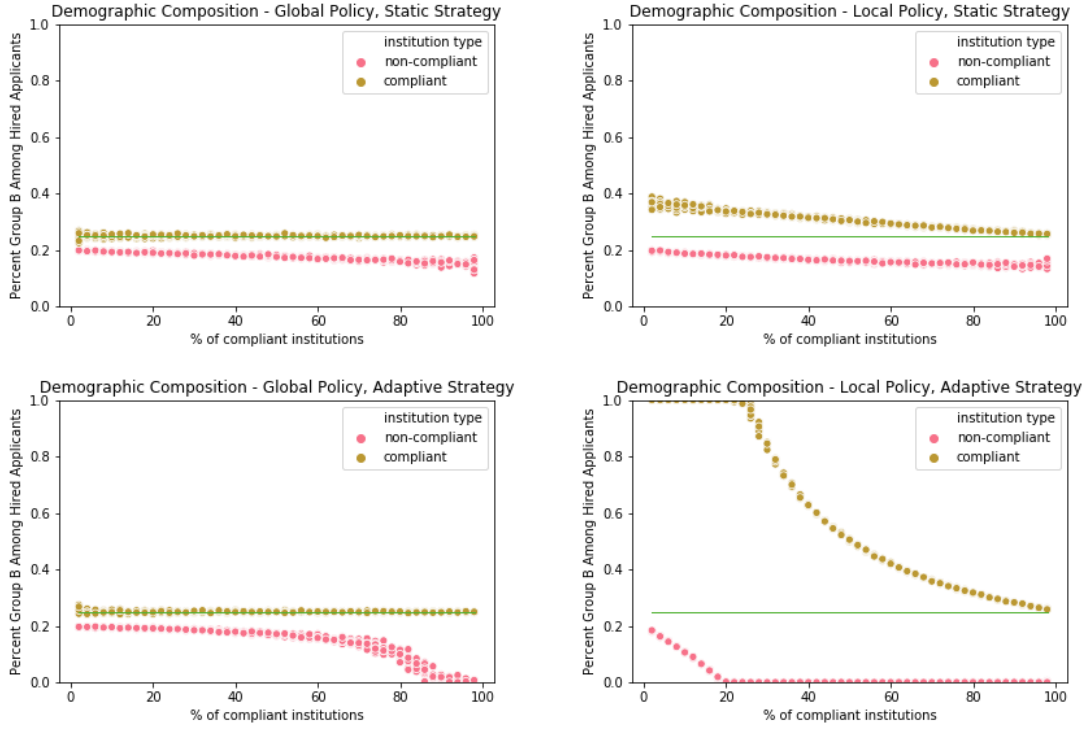


Figure 4: Demographic composition among hired employees, by institution type—in these graphs, Group B is 25% of the population. Top row: Applicants employ a *static* strategy. Bottom row: applicants employ an *adaptive* strategy. Light green horizontal line indicates percentage of Group B in population.

contexts, diversity within the workforce is intrinsically valuable, both due to its potential to enhance team performance and on moral and political grounds [51, 43]. While recent work in fair ML has begun to consider the interaction between diversity, utility and fairness [13, 19], most analyses remain restricted to static settings, focused on individual decision-makers, neglecting the interactions among their decisions and those of their peers and the influence of dynamic factors, such as incentive effects, on long-term policy consequences. Consider what Steel et al. [50] refer to as the representative concept of diversity (see also Smith-Doerr et al. [48]), motivated by concerns about democratic legitimacy, which requires the distributional properties of the selected group to match those of the general population. The global demographic parity measure thus tracks this notion of diversity. Viewed through a static lens, and setting aside the influence of incentives on the choice behavior of applicants, the same connection could be said to hold between the diversity concept and local demographic parity measures. Indeed, this has led some authors to roughly equate these notions of diversity and fairness [13]. The situation becomes more complicated, however, once the dynamics of adaptive application are taken into account. Here, the appearance of (ostensibly desirable) parity at the aggregate level conceals the detrimental impact of local parity policies on diversity within the workforces of the individual employers. These outcomes can emerge absent any explicit desire for segregation on the part of applicants or employers; rather, they are a consequence of the dynamics of incentive effects under partial compliance. In addition to stripping institutions of the benefits

of diversity, the resulting segregation can exacerbate the homophily-based processes that, according to a number of authors [3, 42], can cultivate or amplify injustice.

The aims and the value-alignment of regulation The above discussion indicates the urgent need to clarify the aims and value orientation of regulation. It is useful to frame this issue by inquiring about the aims of the policy that might support the enforcement of local (vs global) demographic parity. In practice, demographic parity is popular, perhaps owing to the 80% rule, and is sometimes invoked as a statistical test in the first phase of disparate impact cases [23]. Note, however, that this connection does not provide a blind endorsement of this form of parity as that which ought to be *enforced*. Certainly, demographic parity can be a part of a *diagnostic toolbox*, serving to indicate disparities that *could*, but need not, indicate underlying discrimination [7, 35]. When precisely measured, demographic disparity can signal moral or legal failings with that particular employer which lie outside the narrow scope of the quantitative measure itself. However, even when the disparity is a symptom of underlying ethical troubles with an allocation policy, partial compliance with the measure may be a misguided remedy (e.g., when the trouble lies with the choice of target outcomes or labels).

Another way of motivating the enforcement (of some form of) demographic parity is by reference to an employer’s wish to implement *affirmative action*. That is, employers may wish to enforce demographic parity, and so preferentially select applicants on the basis of their group membership, as a means of complying with a moral obligation to increase the representation of historically disadvantaged social groups in their institutions. This interpretation resonates with the suggestions that, in some cases, the use of measures such as demographic parity is motivated by the “long-term societal goal” of living in a society where protected attributes are independent of task-relevant outcomes [8]. However, specifying the relation between demographic parity and affirmative action requires clarity about the underlying aim and justifications of the latter—issues that vary radically across different models of affirmative actions [3]—and considerations of whether the former indeed serves those aims. Crucially, our results indicate that, even with minimal incorporation of deployment dynamics, the (partial) adoption of local demographic parity is inconsistent with prominent *future-oriented* justifications of affirmative action. In particular, the emergence of between-institute segregation and a lack of within-institute diversity in our simulations indicate that partial compliance with the measure can result in significant conflicts with diversity-based [24] and integration-based [3] arguments for affirmative action.

Of course, one could adopt a different model of affirmative action to motivate the enforcement of demographic parity. For instance, depending on the interpretation of scores in our model (e.g., as a result of past, upstream injustices, or as an outcome of ongoing biases in an employer’s hiring practices), the measure could be connected to compensation-based (e.g., Thomson [52]) or discrimination-offsetting (e.g., Warren [55]) justifications. Each of these models faces its own set of challenges, including discordance with the actual practice of law, failure to account for the *weight* given to social categories in preferential selection, engendering the expressive harm of *stigmatization*, and undermining the societal legitimacy of affirmative action (see Anderson [3], Fullinwider [24]).

While adjudicating between different models of affirmative action is beyond the scope of this paper, it raises an important concern: Decisions about the aims and the alignment of regulation are value-laden to their core. As a result, these decisions should be made transparently, and on the basis of an integrated consideration of the relevant moral and political models. Importantly, our results show that individual efforts (or the lack thereof) to promote fairness can remain out of sight unless assessed through a more comprehensive, dynamic lens. Analysis of the kind carried out in this paper can not only bring these value

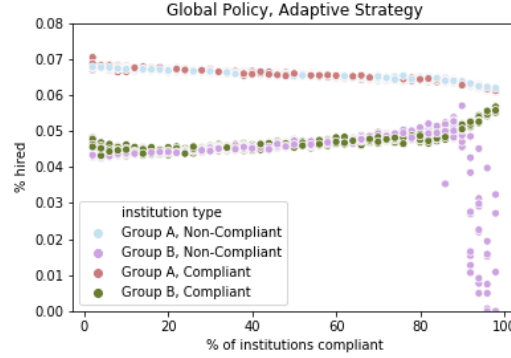


Figure 5: Percent hired per group per institution type. Group B is 25% of the overall population.

judgments into the open, but also complement theorizing about which moral and political models we should prefer, and why.

Partial compliance and the design of appropriate regulatory frameworks The type of partial compliance explored in this paper is a simple representation of the kinds of heterogeneity that exist in the adoption of fairness-promoting measures among various employers both in the use of algorithmic tools in hiring [45] and in hiring more generally. The varied choices of measures is a *consequence* of the ambiguity of current regulatory frameworks. Indeed, the laxity of constraints provides even the *non-compliant* employers in our simulations with a claim to fairness (e.g., by reference to positive predictive parity [18]). Similar to the evaluative practices that inform them, these regulatory frameworks appear to be based on unrealistic assumptions and abstractions of the problem domain.

Our exploration of the dynamics of partial compliance raises central concerns that should inform judgments about both the need for regulation and the form that it should take. For example, in addition to highlighting the potential cost to diversity and integration, our analysis shows that fairness statistics reported by each employer are impacted not only by their policies, but also by those of their competitors. For example, at equilibrium under partial compliance, when employers adopt the global parity policy, an auditor looking at the fractions of applicants from Group A and Group B hired might erroneously conclude that compliant and non-compliant employers were behaving similarly (Figure 5). However, this mistaken view fails to account for the incentive effects, whereby compliant employers come to receive many more applications from members of the disadvantaged group. Thus, when auditing performance vis-a-vis ethical desiderata, we may not be able to determine how a firm is performing without also evaluating their peers. Indeed, in some sense, the abstractions in our model underestimate the implications of partial compliance for current regulatory and evaluative practices. This is because our model represents partial compliance only with respect to *concurrent* policies in a competitive marketplace of hiring. That is, we do not consider allocations that are *upstream* (e.g., in education) and *downstream* (e.g., promotion, mobility across work sector, banking) from hiring decisions, each made by decision-makers who may or may not adhere to their legal (or moral) obligations. Elster [20] makes vivid the significance of such allocations for the well-being and opportunities of individuals:

The life chances of the citizen in modern societies ... depend on allocations made by relatively

autonomous institutions, beginning with admission or nonadmission to nursery school and ending with admission or nonadmission to nursing homes. One could write the fictional biography of a typical citizen, to depict his life as shaped by successive encounters with institutions that have the power to accord or deny him the scarce goods that he seeks [1992].

Despite the potential of unexpected outcome due to robust couplings between policies at successive allocative settings, the implications of partial compliance at successive stages remain under-investigated by the fair ML community. This is a challenge that requires a concerted interdisciplinary effort by the community. Indeed, providing practical guidance under partial compliance poses a challenge to traditional frameworks of distributive justice in political philosophy. While looking to these frameworks for robust conceptual underpinnings of fairness measures can be fruitful [11], they were mainly concerned with modeling the re-distributive obligations of a nation state towards its citizens from the perspective of economic justice. However, when our focus is to provide guidance to relatively autonomous decision-makers using ML tools in local allocative settings, we can no longer simply operate with the same assumptions. Responsible innovation in general [26] and ethical deployment of algorithmic-based decision-making in particular [21] require more comprehensive foresight studies that are equipped to deal with the complexities of deployment context. We hope that our work contributes a few preliminary steps towards this aim.

References

- [1] Dennis J Aigner and Glen G Cain. Statistical theories of discrimination in labor markets. *Ilr Review*, 30(2):175–187, 1977.
- [2] Andrew Altman. Discrimination. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2020 edition, 2020.
- [3] Elizabeth Anderson. *The Imperative of Integration*. Princeton University Press, Princeton, 2010.
- [4] Kwame Anthony Appiah. *As If: Idealization and Ideals*. Harvard University Press, Cambridge, 2017. ISBN 9780674975002.
- [5] Kenneth Arrow et al. The theory of discrimination. *Discrimination in labor markets*, 3(10):3–33, 1973.
- [6] Orley Ashenfelter and Robert S Smith. Compliance with the minimum wage law. *Journal of Political Economy*, 87(2):333–350, 1979.
- [7] Solon Barocas and Andrew D Selbst. Big data’s disparate impact. *Calif. L. Rev.*, 104:671, 2016.
- [8] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. <http://www.fairmlbook.org>.
- [9] Gary S Becker. The economics of discrimination chicago. *University of Chicago*, 1957.
- [10] Federico Bianchi and Flaminio Squazzoni. Agent-based models in sociology. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(4):284–306, 2015.
- [11] Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*, pages 149–159, 2018.
- [12] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building Classifiers with Independency Constraints. In *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops, ICDMW ’09*, pages 13–18, Washington, DC, USA, 2009. IEEE Computer Society.
- [13] L Elisa Celis, Amit Deshpande, Tarun Kathuria, and Nisheeth K Vishnoi. How to be fair and diverse? *arXiv preprint arXiv:1610.07183*, 2016.
- [14] Damon Centola. *How behavior spreads: The science of complex contagions*, volume 3. Princeton University Press, 2018.
- [15] Yang-Ming Chang, Bhavneet Walia, et al. Wage discrimination and partial compliance with the minimum wage law. *Economics Bulletin*, 10(4):1–7, 2007.
- [16] Stephen Coate and Glenn C Loury. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, pages 1220–1240, 1993.
- [17] Alexander D’Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D Sculley, and Yoni Halpern. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 525–534, 2020.
- [18] William Dieterich, Christina Mendoza, and Tim Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 2016.
- [19] Marina Drosou, HV Jagadish, Evaggelia Pitoura, and Julia Stoyanovich. Diversity in big data: A review. *Big data*, 5(2):73–84, 2017.

- [20] Jon Elster. *Local justice: How institutions allocate scarce goods and necessary burdens*. Russell Sage Foundation, 1992.
- [21] Sina Fazelpour and Zachary C Lipton. Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 57–63, 2020.
- [22] 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*, KDD ’15, pages 259–268, New York, NY, USA, 2015. ACM.
- [23] 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.
- [24] Robert Fullinwider. Affirmative Action. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2018 edition, 2018.
- [25] Ben Green and Lily Hu. The myth in the methodology: Towards a recontextualization of fairness in machine learning. In *Proceedings of the machine learning: the debates workshop*, 2018.
- [26] Armin Grunwald. Technology assessment for responsible innovation. In *Responsible Innovation 1*, pages 15–31. Springer, 2014.
- [27] S Hansson. *The ethics of risk: Ethical analysis in an uncertain world*. Springer, 2013.
- [28] 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.
- [29] Hoda Heidari, Vedant Nanda, and Krishna P Gummadi. On the long-term impact of algorithmic decision policies: Effort unfairness and feature segregation through social learning. *arXiv preprint arXiv:1903.01209*, 2019.
- [30] 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.
- [31] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Conference on Fairness Accountability and Transparency (FAT*)*, 2018.
- [32] Maximilian Kasy and Rediet Abebe. Fairness, equality, and power in algorithmic decision making. Technical report, Working paper, 2020.
- [33] Elmar Kiesling, Markus Günther, Christian Stummer, and Lea M Wakolbinger. Agent-based simulation of innovation diffusion: a review. *Central European Journal of Operations Research*, 20(2):183–230, 2012.
- [34] Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? In *ACM Conference on Economics and Computation (EC)*, 2019.
- [35] Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. Does mitigating ml’s impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems*, pages 8125–8135, 2018.

- [36] Zachary C. Lipton and Jacob Steinhardt. Troubling Trends in Machine Learning Scholarship. *Communications of the ACM (CACM)*, 62(6):45–53, 2018.
- [37] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 6196–6200. AAAI Press, 2019.
- [38] Lydia T Liu, Ashia Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs, and Jennifer Chayes. The disparate equilibria of algorithmic decision making when individuals invest rationally. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 381–391, 2020.
- [39] David Miller. Taking up the slack? responsibility and justice in situations of partial compliance. In Carl Knight and Zofia Stemplowska, editors, *Responsibility and Distributive Justice*, pages 230–45. Oxford University Press, 2011.
- [40] Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. The social cost of strategic classification. In *Conference on Fairness Accountability and Transparency (FAT*)*, 2018.
- [41] Ryan Muldoon. *Social contract theory for a diverse world: Beyond tolerance*. Taylor & Francis, 2016.
- [42] Cailin O’Connor. *The Origins of Unfairness*. Oxford University Press, Oxford, 2019.
- [43] Scott E Page. *The diversity bonus: How great teams pay off in the knowledge economy*. Princeton University Press, 2019.
- [44] Edmund S Phelps. The statistical theory of racism and sexism. *The american economic review*, 62(4): 659–661, 1972.
- [45] Javier Sánchez-Monedero, Lina Dencik, and Lilian Edwards. What does it mean to ‘solve’ the problem of discrimination in hiring? social, technical and legal perspectives from the uk on automated hiring systems. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 458–468, 2020.
- [46] Tamar Schapiro. Compliance, complicity, and the nature of nonideal conditions. *The Journal of Philosophy*, 2003.
- [47] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Fairness, Accountability, and Transparency (FAT*)*, 2019.
- [48] Laurel Smith-Doerr, Sharla N Alegria, and Timothy Sacco. How diversity matters in the us science and engineering workforce: A critical review considering integration in teams, fields, and organizational contexts. *Engaging Science, Technology, and Society*, 3:139–153, 2017.
- [49] Lyn Squire and Sethaput Suthiwart-Narueput. The impact of labor market regulations. *The World Bank Economic Review*, pages 119–143, 1997.
- [50] Daniel Steel, Sina Fazelpour, Kinley Gillette, Bianca Crewe, and Michael Burgess. Multiple diversity concepts and their ethical-epistemic implications. *European Journal for Philosophy of Science*, 8(3): 761–780, 2018.
- [51] Daniel Steel, Sina Fazelpour, Bianca Crewe, and Kinley Gillette. Information elaboration and epistemic effects of diversity. *Synthese*, pages 1–21, 2019.

- [52] Judith Jarvis Thomson. Preferential hiring. *Philosophy & Public Affairs*, pages 364–384, 1973.
- [53] Laura Valentini. Ideal vs. Non-ideal Theory: A Conceptual Map. *Philosophy Compass*, 2012.
- [54] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai. *Available at SSRN*, 2020.
- [55] Mary Anne Warren. Secondary sexism and quota hiring. *Philosophy & Public Affairs*, pages 240–261, 1977.
- [56] Kevin JS Zollman. Network epistemology: Communication in epistemic communities. *Philosophy Compass*, 8(1):15–27, 2013.