Dear Reviewers and Meta-Reviewers,

In the attached, revised manuscript, please note the changes described herein that we made to address the sociotechnical issues raised by our original submission. We believe we have now better clarified our results to avoid potential misinterpretations. We have highlighted in blue the sections of the manuscript which have been substantively revised to address these concerns.

- Our revisions largely concentrate on the introduction (Section 1) and discussions (Section 5) to clarify our framing and address points of potential misunderstandings.
- We added a concrete running example that we use to explain our assumptions and modelling choices in Section 2. Our notation and mathematical formulation are unchanged.
- Section 3 & 4 are our main technical sections - we have only added explanations in connection to our running example. Section 4 also now features new paragraph that explains the differences between our intervention and the existing mechanisms.

Our problem formulation and technical results are preserved and remain mostly untouched; to the unhighlighted sections of the document, we have made only minor adjustments to phrasing, presentation, or organization to improve clarity and consistency. We thank you, and our reviewers, for their feedback, comments, and concerns; we believe this process has resulted in a stronger presentation and much more meaningful contribution to the field.

Sincerely,
Authors of Paper 11361

We restate (in *italicized text*) and address the specific changes requested by the sociotechnical review sequentially.

1. *Review and update analogies drawn from evolutionary biology, so as to minimize the risk of essentialist misinterpretation.*

The connection between replicator dynamics and biological evolution is now delayed until it can be appropriately qualified—Direct reference to "evolution" is now absent from both our title and our abstract. In the same paragraph in which the connection first appears (line 63), we emphasize that the phenomenon applies to "competing *strategies for qualification* (which determine true machine learning labels $\{0, 1\}$, affecting agent utilities) [which] replicate within groups (*i.e.*, isolated subpopulations that differ only in size and initial proportions of qualified individuals)" (lines 64–66) and that "only the *profile of strategies* in each subpopulation is subject to evolution—narrowly qualified by the competition between strategies for replicative success—rather than the subpopulations themselves" (lines 72–74). This connection is again carefully qualified in lines (202–207), where fitness is defined, as "notably, the fitness of a strategy depends only on the classifier—not group membership G. Agents remain identically modelled across all groups."

We repeatedly highlight throughout the revised manuscript that we model groups as equivalent in all respects except for initial qualification rate (and possibly size), making this assertion in the abstract (lines 17–18); introduction (lines 65–66), (line 69), (lines 110–112); and formulation (lines 158–160), (line 207). Moreover, we explicitly address the equity of such assumptions in our introduction (lines 48–50).

Finally, the citation of Zohary's work on unconscious plant domestication, appearing in the original manuscript, has also been removed due to the potential of gross misunderstanding of our intended analogy to the modelled *process* explored by our work. We wish to avoid analogies regarding the *outcome* of real-world discrimination and plant domestication.

2. *Connect technical definitions to some real-world examples, and discuss how the assumptions made in the paper align or not with those real-world examples. Some of the key terms that require connecting to examples are "qualification" and "subpopulation". The best approach would be to introduce a plausible running example that readers can use to understand the*

implications of the results. Such an example would need to pin down what precisely is meant by "qualification", how the groups (subpopulations) are defined, and what the classifier is doing. If no plausible example can be produced, then the authors need to walk back their statements about the applicability of the results to real-world sociotechnical systems.

In the revised manuscript, we now appeal to a running example using loan applications, in which "a regional bank (classifier) serving several isolated communities (groups, subpopulations) [offers] standardized loans for which every individual (agent) applies"(Formulation: lines 125–127). We use it to ground interpretation of *qualification* as ex post repayment of a loan if approved (lines 134–136) and provide examples of *strategies* such building credit or seeking a wage increase. Additional connections to this specific example occur in the abstract (lines 5–10); introduction (lines 33–34, 66-69); formulation (lines 139, 146–147, 154–157, 169–171, 182–191); discussion of dynamics (lines 227–229, 278–280); and discussion of interventions (lines 312-313, 331, 362–363), thus motivating our models of agent and classifier actions as well as our analytical statements (*e.g.*, regarding stability). We believe our appeal to this extended example, as well as a brief reference to other possible examples such as "hiring decisions [citing Coate and Loury] or college admissions", ensures that our work can be connected to the real world.

We also take greater care in interpreting *qualification*, *groups/subpopulations*, *fairness* (see point 3 herein), and *disparity* in general, ensuring that our use of such terms is clear. In particular, we ensure that qualification is understood as a binary label with an ex post, positive interpretation (line 68, 134–135), mutable via imitation of successful strategies (line 9, 42, *etc.*) (see point 4 herein) and that groups are understood in terms of information flow and imitation, rather than "sensitive attributes" (lines 50–55, 212–218).

3. *Be explicit about the operationalization of fairness and explain what the various assumptions and criteria around that operationalization mean in the context of real-world dynamics.*

The introduction is now explicit about fairness operationalizations, explicitly differentiating "long-term causal fairness" and "normative present fairness", which we directly reference in our discussion of different interventions (lines 354–370). Moreover, we tie such interventions to our running example of loan applications to ground their interpretation. (lines 312–313, 331).

4. *Discuss how the replicator dynamics might be realized in real-world examples.*

Our appeal to mechanisms that plausibly justify the use of replicator dynamics (*i.e.*, the imitation of strategies) is now mathematically specific (Section 2.2: lines 192–193, footnote 4 on page 5), and we reference prior work that derives replicator dynamics from these mechanisms (lines 116–120, lines 192–194). More importantly, we justify the plausible realization of such mechanisms in terms of our running example (lines 182–191).

5. *Connect the proposed intervention to a clear motivation of mitigating harms to marginalized groups, and more explicitly clarify differences between this intervention and other approaches (such as equalized odds and demographic parity) in service of that overall goal.*

Our exposition now explictly discusses interventions in terms of different types of fairness (see point 3 herein) and connects to our running example (lines 354–370). We seek to mitigate harm by "[reforming] the misapplication of machine learning, inequitable modelling, and myopic notions of fairness" (lines 393–394). In particular, we explicitly qualify our goals of developing "an *equitable* model of population response: one which does not suppose fixed differences between groups of people but endogenizes disparities by appeal to unequal *initial conditions* and accounts for any group-specific environmental conditions as dynamical variables. We reform our notion of "groups" (*i.e.*, *subpopulations*) to appeal to natural boundaries of information exchange rather than artificially imposed classes of people. We thus offer a potentially more meaningful way to group individuals in discussions of fairness, asserting that, when considering such networks of peer exchange, "sensitive attributes" such as race, sex, color, *etc.* might not correspond to meaningful divisions of people, which depend on social context." (lines 48-58)

Unconscious Selection: Persistent Qualification Rate Disparities and Interventions

Anonymous Author(s)
Affiliation
Address
email

Abstract

Realistically—and equitably—modeling the dynamics of group-level disparities in machine learning remains an open problem. In particular, we desire models that do *not* suppose inherent differences between artificial groups of people—but rather endogenize disparities by appeal to unequal *initial conditions* of insular subpopulations. In this paper, agents each have a real-valued feature X (*e.g.*, credit score) informed by a "true" binary label Y representing *qualification* (*e.g.*, for a loan). Each agent alternately (1) receives a binary classification label $\hat{Y}$ (*e.g.*, loan approval) from a Bayes-optimal machine learning classifier observing X and (2) may update their qualification Y by imitating successful *strategies* (*e.g.*, seek a raise) within an isolated group G of agents to which they belong. We consider the disparity of qualification rates $\Pr(Y = 1)$ between different groups and how this disparity changes subject to a sequence of Bayes-optimal classifiers repeatedly retrained on the global population. We model the evolving qualification rates of each subpopulation (group) using the replicator equation, which derives from a class of imitation processes. We show that differences in qualification rates between subpopulations (groups) can persist indefinitely for a set of non-trivial equilibrium states due to deployments of unaware classifiers, even when groups are identical in all aspects except initial qualification densities. We next simulate the effects of commonly proposed fairness interventions on this dynamical system along with a new feedback control mechanism capable of permanently eliminating group-level qualification rate disparities. We conclude by discussing the limitations of our model and findings and by outlining potential future work.

1 Introduction

Algorithmic prediction is increasingly used for socially consequential decisions and may determine individual access to information, education, employment, credit, housing, medical treatment, freedom from incarceration, or freedom from military targeting [1–5]. This situation raises technical challenges and ethical concerns, particularly regarding the dynamics of systemic inequalities and attendant harms to society [6–8]. Nonetheless, realistically—and equitably—modeling the dynamics of disparity in machine learning remains an open problem.

Research historically considered fairness as the statistical regularity of algorithmic predictions [9] (*e.g.*, across groups [10–16], or between individuals [10, 11]) or appealed to preference guarantees [17–19] or causal considerations [19, 20], but ignored the *response* of a population to new prediction policies. For example, the rate at which potential loan applicants seek higher wages, falsify income, or forego application might change if banks use new credit models to approve or deny loans. We refer this class of fairness definitions as *normative present fairness*.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Efforts to model population response [21–30] and the autonomous dynamical systems arising from mutual recursion with myopically updating prediction policies [21–28] have expanded, but it has remained to plausibly explain persistent disparities under group-independent prediction policies—*i.e.*, those that do not discriminate on the basis of group membership—without making structurally unfair assumptions. Our paper contributes to this literature and considers the long-term consequences of machine learning on inter-group disparities when a sequence of classifiers induces dynamics within the rates of strategy adoption in each group. Upon adopting a dynamical framework, we note that yet another operationalization of fairness arises: the asymptotic equality of latent variables (*i.e.*, those causally responsible for outcome disparities) between groups. This notion of *long-term fairness* need not be consistent with *normative present fairness*, which may actively combat it, highlighting a tension between *ends* and *means* for fairness considerations.

1.1 Our Contributions

Herein, we describe an *equitable* model of population response: one which does not suppose fixed differences between groups of people but endogenizes disparities by appeal to unequal *initial conditions* and accounts for any group-specific environmental conditions as dynamical variables. We reform our notion of "groups" (*i.e.*, *subpopulations*) to appeal to natural boundaries of information exchange rather than artificially imposed classes of people. We thus offer a potentially more meaningful way to group individuals in discussions of fairness, asserting that, when considering such networks of peer exchange, "sensitive attributes" such as race, sex, color, *etc.* might not correspond to meaningful divisions of people, which depend on social context. Finally, we recognize fairness interventions as dynamical control policies that (un)consciously select the future trajectories of a given system. We are therefore allowed to consider interventions that explicitly incorporate *feedback* from dynamical variables—rather than relying on fixed, prescriptive modifications of predictor loss functions.

Our first contribution is to propose a model of *classifier-induced* group-level strategy adoption that is (1) equitable, *i.e.*, free from structurally unfair assumptions as described above, (2) capable of explaining persistent disparities under Bayes-optimal, group-independent policies, and (3) derivable from plausible, localized information exchange between individuals. Specifically, we appeal to the replicator equation, an established model for evolutionary phenomena without mutation, to model how competing *strategies for qualification* (which determine true machine learning labels $\{0, 1\}$, affecting agent utilities) replicate within groups (*i.e.*, isolated subpopulations that differ only in size and initial proportions of qualified individuals). We ground statements with a running example involving loan applications (elaborated upon in Section 2.2) for which qualification (label $Y = 1$), interpreted as being in the public interest, implies future repayment of a loan for an applicant with feature profile X. As we avoid assuming inherent differences between groups, we consider the label-conditioned feature distribution $\Pr(X \mid Y = y)$ as group-independent and define qualification disparity in terms of differences in group qualification rates $\Pr(Y = 1)$. We formulate our model in Section 2, emphasizing that only the *profile of strategies* in each subpopulation is subject to evolution—narrowly qualified by the competition between strategies for replicative success—rather than the subpopulations themselves. The persistence of disparity is thus attributed to classifier policy.

Our second contribution, in Section 3, is a rigorous examination of the dynamical system formed by the replicator equation and an updating, group-independent, Bayes-optimal classifier policy, including a characterization of its equilibrium states with linear stability analysis. We identify the set of stable interior states of the system as a stable hyperplane and show that any initial state with non-zero total qualification disparity, defined in Section 3, will continue to exhibit non-zero disparity asymptotically if the state attracts to the stable hyperplane (Theorem 10). In this sense, we claim that qualification rate disparity persists indefinitely for this setting.

Our final contribution, in Section 4 is to consider a dynamics-aware fairness intervention based on feedback control that parametrically violates classifier group-independence (and therefore, in our setting, *equalized odds* [12–14] and *envy-freeness* [17, 18]) to achieve long-term fairness. We use simulation to contrast this feedback control policy to a group-independent classifier, a policy subject to demographic parity [10, 11], and "laissez-faire", *group-specific* policies. We conclude by discussing the limitations of our model and our findings and by outlining potential future work.

1.2 Related Work

Our work chiefly contributes to the literature on fairness in machine learning, but also builds on prior work on "statistical discrimination". The most relevant publications are those which have highlighted

the importance of studying the dynamics and long-term consequences of machine learning, fairness constraints, and models of population response. In particular, Liu et al. [29] used Markov transitions to model population response in terms of features without considering classifier retraining; D'Amour et al. [22] and Zhang et al. [23]. used Markov transitions for agent labels with classifier retraining; Zhang et al. [31] modelled agents' decisions of whether to engage with classification, but limited changes to group-conditioned distributions of features or labels to be "relatively insensitive" to changes in classifier policy; Coate and Loury [21], Hu and Chen [27], and Liu et al. [26] considered economical "best-response" models in agent labels with classifier updates; and Heidari et al. [24] considered an imitation-based model of social learning in which agents choose between the strategies of other agents to maximize utility and minimize effort. Tang et al. [32] also studied the delayed and accumulated impacts of past deployed policies, but did not study the fairness implication of such impacts. Similarly, literature on "fair bandit/reinforcement learning" [33–35] has largely focused on technical aspects of imposing normative present fairness in a sequential setting.

Our proposed model synthesizes prior conceptual innovations: First, Coate and Loury [21], Hu and Chen [27], and Ensign et al. [36] each considered incomplete information available to a classifier as a means to equitably endogenize persistent predictor bias, but did not consider incomplete information available to individual agents. Second, the class of response functions considered by Mouzannar et al. [28] allows group-level strategic responses to depend on existing qualification rates and may be used to endogenize persistent disparity under group-independent policies. The cited work does not explore this direction, but, like us, the authors assume "groups are ex-ante equal in all respect except for their qualification profiles...and any potential coupling between groups can only happen through the different and interacting selection rates induced by the policies"[28, p. 362]. Above this, we cite a plausible mechanism of imitation, motivated by incomplete information available to individual agents, to justify replicator dynamics as a special case of the proposed response functions, and we extend a dynamical analysis for a classifier forced to contend with misclassification errors.

In supporting the use of the replicator equation to model aggregated, group-level decisions, we cite the imitation-based derivation(s) the replicator equation by Björnerstedt and Weibull [37]; the characterization of evolutionarily stable strategies conducted by Taylor and Jonker [38]; the analogy of memes as attributed to Dawkins [39]; and the extensive application of the replicator equation in game-theoretic contexts as explored by Friedman and Sinervo [40].

2 Formulation

Note that we defer all proofs to the supplementary material.

We consider countably many *agents*, $n \geq 2$ *groups*, and a single *classifier*. Until Section 2.2, our setting matches that of Coate and Loury [21] but treats n groups and more a granular classifier utility function. We ground statements with a running example: a regional bank (classifier) serving several isolated communities (groups, subpopulations) by offering standardized loans for which every individual (agent) applies. Other examples include hiring decisions [21] or college admissions.

Agents belong to **groups**, interpreted in Section 2.2 and consistent with *isolated communities* in our running example, with known relative frequencies $\mu_g \in (0, 1)$. We vectorize these frequencies as $\boldsymbol{\mu}$.

$$\mu_g \coloneqq \Pr(G = g); \quad \forall g.\ \mu_g \in (0,1); \quad \sum_{g \in \mathcal{G}} \mu_g = 1; \quad \boldsymbol{\mu} \coloneqq (\mu_1, \mu_2, ..., \mu_n) \tag{1}$$

For all statements of probability, we assign uniform probability mass to each agent.

In addition to relative size μ_g, each group has a **qualification rate** $s_g \in (0, 1)$, which we vectorize as our **state** variable $\mathbf{s}$. We denote the global qualification rate as $\overline{s}$:

$$s_g \coloneqq \Pr(Y = 1 \mid G = g); \quad \mathbf{s} \coloneqq (s_1, s_2, ..., s_n); \quad \overline{s} \coloneqq \sum_{g \in \mathcal{G}} \mu_g s_g = \langle \boldsymbol{\mu}, \mathbf{s} \rangle \tag{2}$$

Assumption 1. *No community is completely (un)qualified.* $\forall g.\ s_g \in (0, 1)$

In our banking example, a qualified ($Y = 1$) individual will pay back a loan in full if accepted ($\hat{Y} = 1$), an outcome we declare extrinsically desirable. The fraction of qualified individuals in community g is s_g, and the fraction of qualified individuals in the total population is $\overline{s}$. Assumption 1 states no community is completely (un)qualified. Knowing $\mu_g \in (0, 1)$, nor is the total population.

Table 1: Agent-specific variables forming a Markov chain

Variable	Meaning	Domain	Realizations
G	group	$\mathcal{G} = \{1, 2, ..., n\}$	g, h, i, j
Y	qualification	$\{0, 1\}$ *i.e.*, {unqualified, qualified}	y
X	feature	$(-\infty, \infty)$	x
$\hat{Y}$	classification	$\{0, 1\}$ *i.e.*, {reject, accept}	$\hat{y}$

The **feature** X of an agent qualified as $(Y = y)$ is sampled according to a probability density function q_y. In our banking example, we may interpret X as a "credit score" known to the bank.

$$q_y(x) \coloneqq p_X(x \mid Y = y); \quad y \in \{0, 1\} \tag{3}$$

Assumption 2. *The qualification-conditioned distribution of features $q_y(x)$ is group-independent.*

Assumption 2 ensures that qualified individuals are statistically indistinguishable in terms of feature X across different communities—as are unqualified individuals. Given an agent's qualification Y, learning G gives no additional information about X.

Assumption 3. *q_y is differentiable and strictly positive for each y. The values of X are ordered and unified such that $q_1(x)/q_0(x)$ is strictly increasing in X:*

$$\forall x, y.\ q_y(x) \in (0, \infty); \quad \frac{d}{dx}\left(\frac{q_1(x)}{q_0(x)}\right) > 0 \tag{4}$$

Assumption 3 ensures that the feature X is "well-behaved". In our example, as credit scores increase as x, the odds that individuals with that credit score x will pay off loans also increases.

Finally, a **classifier** observes the feature X of each agent, from which it must predict the agent's correct label Y using a deterministic policy π, ignorant of G.

Assumption 4. *The classifier learns the true distribution* $\Pr(Y \mid X)$ [1].

Assumption 5. *The classifier maximizes expected utility u with risk-neutral preferences. This utility u is linear in each outcome fraction $\Pr(Y = y, \hat{Y} = \hat{y})$, with coefficients [2] $V_{y,\hat{y}} \in (-\infty, \infty)$ that are independent of feature value X and group membership G. The classifier prefers correct predictions.*

$$\hat{Y} \coloneqq \pi(X); \quad u(\pi) \coloneqq \sum_{y,\hat{y}=0}^{1} V_{y,\hat{y}} \Pr(Y = y, \pi(X) = \hat{y}); \quad V_{y=\hat{y}} > V_{y\neq\hat{y}} \tag{5}$$

In our example, Assumption 5 suggests that the bank maximizes expected net profit, and that the bank expects net profit proportional to $V_{y,\hat{y}}$ from each individual qualified as y and approved as $\hat{y}$, independent of credit score or community G. By Assumption 4, the bank selects policy π knowing the stochastic relationship between qualification Y and credit score X for the region it serves.

With group-independent classifier policies, having excised assumptions of inherent differences between groups in our formulation, we emphasize that unequal group qualification rates are wholly responsible for outcome disparities. We consider eliminating differences in group qualification rates as a realization of long-term fairness.

Theorem 1. *Discounting sets of measure zero, the u-maximizing, group-independent policy π is parameterized by the **feature threshold** $\phi \in [-\infty, \infty]$ such that $\pi(x) = 1$ if and only if $x > \phi$, where ϕ depends only on the global qualification rate $\bar{s}$ as*

$$\xi \coloneqq \frac{V_{0\hat{0}} - V_{0\hat{1}}}{V_{1\hat{1}} - V_{1\hat{0}}}; \quad \frac{q_1(\phi)}{q_0(\phi)} = \xi\left(\frac{1 - \bar{s}}{\bar{s}}\right) \tag{6}$$

*When a solution in ϕ to the **threshold equation**, Eq. (6), does not exist, ϕ is either $\pm\infty$.*

Corollary 1.1. *The classifier's feature threshold ϕ responds inversely to $\bar{s}$: $\frac{d\phi}{d\bar{s}} < 0, \quad \frac{d\bar{s}}{d\phi} < 0$.*

[1] In practice, this distribution may be learned from sufficient data

[2] We will abuse notation to write, *e.g.*, $V_{y,\hat{y}}$ as $V_{1\hat{0}}$, to disambiguate the order of indices on V and, later, U

Assumption 6. *V is such that $\xi \in (0, \infty)$*

The threshold equation, Eq. (6) is a reprise of Coate and Loury [21], restricted to group-independent classifier policies. In our example, the bank maximizes its utility by approving individuals with credit scores greater than ϕ and denying everyone else. Interpreting Assumption 6, there exist populations for which the bank prefers to accept some applicants and reject others.

2.1 Time-dependence

We model our system in discrete time for semantic reasons, acknowledging that a learning process consistent with Assumption 4 requires time, although the mathematics generalize to continuous time without issue [3]. Where required, we will denote time-dependence in square brackets $[t]$. Where we omit this explicit dependence, as for all prior expressions, it is understood that all variables in an expression correspond to the same time t.

Assumption 7. *The relative sizes of groups μ_g, qualification-conditioned feature distributions q_y, and classifier utility coefficients $V_{y,\hat{y}}$ are all time-independent. All prior assumptions hold independently for each time step.*

2.2 Replicator dynamics

Anticipating algorithmic classification, how do agents decide whether to become qualified? In our banking example, we imagine that individuals must "invest" or "apply capital" to be able to pay back loans and that the "rationality" of doing so depends on what they know about the classification policy and potential outcomes—which they must estimate from incomplete information provided by peer examples. To deal with the uncertainty of limited examples, we imagine the emergent heuristic of updating personal qualification by *imitating* the strategies of others based on popularity and "success". For example, if your friend chose to become qualified for a loan and now runs a small business, the success of the business may induce you to seek qualification yourself by first building credit history; if many of your neighbors receive loans despite being unqualified and appear successful investing in speculative assets, you may, likewise, consider qualification to be a waste of resources.

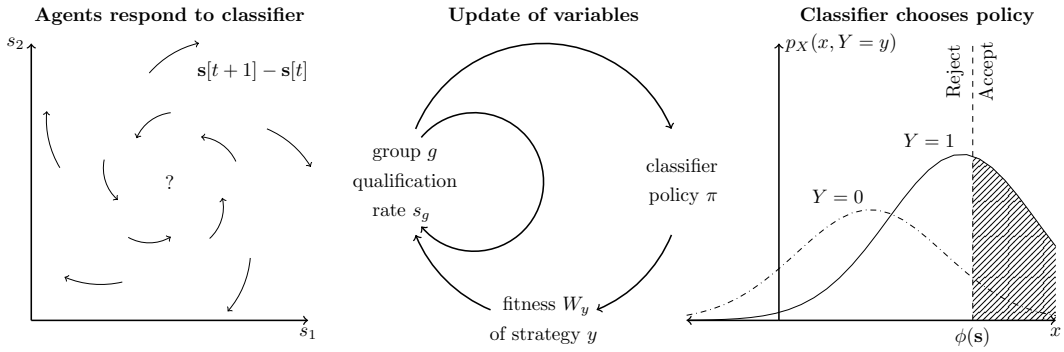

Figure 1: Our model appeals to the replicator equation Eq. (7) to model population response and considers a Bayes-optimal, group-independent classifier policy π with feature threshold ϕ (Eq. (6), right pane). When coupled (middle pane), these equations give rise to an autonomous dynamical system. We wish to understand how the vector of group qualification rates $\mathbf{s}$, as our state variable, changes in time (left pane).

Björnerstedt and Weibull [37] have shown that *imitation* in this form, whereby agents stochastically update to strategies weighted by success and popularity [4] yields the (continuous time) **replicator equation**, which we use in its discrete time form, as detailed by Friedman and Sinervo [40]:

$$s_g[t+1] = s_g[t]\frac{W_1[t]}{\overline{W}_g[t]}; \quad \overline{W}_g := W_1 s_g + W_0(1 - s_g) \tag{7}$$

[3] The continuous replicator equation appears in Björnerstedt and Weibull [37] and Friedman and Sinervo [40]

[4] Weighting by popularity effects "preferential attachment" for the success of a strategy in a subpopulation and thus introduces dynamical inertia. Replicator dynamics also arise when agents update strategies with (Poisson distributed) expected periodicity that is affine in the success of one's current strategy [37].

Here, W_y is the **fitness** of strategy y, which, by Theorem 2, we model as independent of feature X and group G. Following Björnerstedt and Weibull [37], we may derive the fitness W_y in terms of expected "success" $U_{y,\hat{y}}$ of each qualification-classification outcome $(y, \hat{y})$:

Assumption 8. *The fitness of strategy $(Y = y)$, denoted as W_y^g, is affine in the average success $U_{y,\hat{y}}$ of qualification $(Y = y)$ with classification $(\hat{Y} = \hat{y})$. $U_{y,\hat{y}}$ is time-, feature- and group-independent. Without loss of generality, we restrict $U_{y,\hat{y}} \in [0, \infty)$ and drop the constant bias term from each W_y^g.*

$$W_y^g := \sum_{\hat{y}=0}^{1} \Pr(\hat{Y} = \hat{y} \mid Y = y, G = g) U_{y,\hat{y}} \tag{8}$$

Theorem 2. *The fitness W_y^g of strategy $Y = y$ in group g is feature- and group-independent.*

$$W_y^g = W_y = U_{y\hat{1}} + (U_{y\hat{0}} - U_{y\hat{1}}) Q_y(\phi), \quad y \in \{0, 1\} \tag{9}$$

Intuitively, "success" U may be interpreted as *utility* or *payoff* to each agent when agents align strategy adoption with personal incentives, but, fundamentally, W corresponds to the relative success of the *strategy* in replicating, *i.e.*, spreading between individuals. The strategies of (non)qualification are thus subject to evolutionary pressures, competing to out-replicate each other in an environment shaped by perceptions of classifier policy. Notably, the fitness of a strategy depends only on the classifier—not group membership G. Agents remain identically modelled across all groups.

Assumption 9. *The success of (non)qualification is sensitive to classification, and the expected success for qualified individuals increases with classifier acceptance: $U_{0\hat{1}} \neq U_{0\hat{0}}$; $U_{1\hat{1}} > U_{1\hat{0}}$*

Assumption 10. *Each group g has the properties of a closed population in which* qualification*, as a strategy or meme [39], competes with* non-qualification *free from exchange with other groups.*

It is significant that we model population updates as independent for *closed* populations, as this restricts our interpretation of groups, which must be functionally impermeable to the exchange of qualification strategies. To precisely delineate between real-world examples of "groups" is akin to disassociating "cultures", which also imply boundaries of exchange but generally intersect. Having noted that "sensitive attributes" such as race, sex, color, *etc.* may not correspond to meaningful divisions between people (which depend on social context), we instead qualify a **group** by the extent to which it satisfies Assumption 10. As an open question, we ask whether imposing arbitrary demographic-dependent policies may catalyze the formation of groups of strategic peers, but we will consider insular social groups or isolated communities as canonical examples.

3 Dynamics

The threshold equation, Eq. (6): $\phi[t](\mathbf{s}[t])$, and the replicator equation, Eq. (7): $\mathbf{s}[t+1](\phi[t], \mathbf{s}[t])$, may be coupled to yield an autonomous dynamical system $\mathbf{s}[t+1](\mathbf{s}[t])$ that evolves in time. To analyze it, we first generate a useful set of coordinates to compliment $\overline{s}$ and track qualification disparities, defined by the differences in s_g between groups. We then note the importance of $W_1(\phi) - W_0(\phi)$ to the overall dynamics of the system, and use it to identify non-trivial equilibrium states. To interpret this section for our example, we ask how community-specific loan qualification rates change as individuals imitate successful strategies in their isolated communities, assuming the bank maximizes profit using group-independent credit thresholds for loan approval.

Definition 3. Define the (signed) **qualification distance** from group h to group g as

$$\delta(g, h) := s_g - s_h, \quad g, h \in \{1, 2, ..., n\} \tag{10}$$

We next define the vector D comprising $(n-1)$ linearly-independent qualification distances between sequential pairs of subpopulations:

$$D := \Big(\delta(1,2),\ \delta(2,3),\ ...,\ \delta(n-1,n)\Big) \tag{11}$$

The components of D and value of $\overline{s}$ together yield a complete set of coordinates to describe the state of the dynamical system, which we may exchange for the original vector of qualification rates $\mathbf{s} = (s_1, s_2, ..., s_n)$ via a non-orthogonal, linear change of basis:

$$s_g = \overline{s} + \sum_{h=g}^{n-1} \delta(h, h+1) - \sum_{h=1}^{n-1} \sum_{k=1}^{h} \mu_k \delta(h, h+1) \quad \forall g \in \mathcal{G} \tag{12}$$

Let us denote the state vector in our new coordinate system as $\mathbf{r} := \big(\delta(1,2), \delta(2,3), ..., \delta(n-1,n), \overline{s}\big)$ and refer to any p-norm of D, $\|D\|_p$, as the **total qualification rate disparity** of corresponding state.

Remark 4. By linearity of $\overline{s}$ (Eq. (2)), we may describe all states with a specific $\overline{s}$ as a **hyperplane**.

Remark 5. The nullity of total qualification disparity, $\|D[t]\|_p$, where $p \geq 1$, is preserved in time.

$$p \geq 1; \quad \big\|D[t]\big\|_p := \Big(\sum_{g=1}^{n-1} \big|\delta(g, g+1)[t]\big|^p\Big)^{1/p}; \quad \big\|D[t]\big\|_p = 0 \iff \big\|D[t+1]\big\|_p = 0 \tag{13}$$

Remark 5 highlights a weak notion of the persistence of disparity within the system sans intervention: Any state that possesses some non-zero total qualification disparity (defined as some chosen p-norm of D) must always exhibit some non-zero total qualification disparity with any finite time horizon. In our example, if some communities start more qualified than others, the qualification rates of different communities will not naturally equalize in any given lifetime. Note that this statement is insufficient to address the limit $t \to \infty$, however. For a stronger result, Theorem 10, that includes this limit, we must characterize the system's equilibrium states.

3.1 Equilibrium

Definition 6. The system as a whole is **at equilibrium** when, for all $g \in \mathcal{G}$ simultaneously, s_g is stationary in time:

$$\text{at equilibrium} \overset{\text{def}}{\iff} \forall g \in \mathcal{G}.\ \exists t_0 \text{ s.t. } \forall t \geq t_0, \quad s_g[t] = s_g[t_0] \tag{14}$$

Note: replicator dynamics is an instance of the more general family of monotone dynamics, with which all equilibria are shared [37, 40].

Theorem 7. *Disregarding boundary states by Assumption 1, the replicator equation, Eq.* (7)*, implies*

$$\operatorname{sgn}\big(\overline{s}[t+1] - \overline{s}[t]\big) = \operatorname{sgn}\big(W_1(\phi) - W_0(\phi)\big) \tag{15}$$

Theorem 8. *It is necessary and sufficient for a system at equilibrium that $W_1 = W_0$ or for the system to occupy some vertex of the state space.*

$$\text{at equilibrium} \iff \begin{cases} W_1 = W_0 & \textit{(internal equilibrium)} \\ \forall g \in \mathcal{G}. \quad s_g \in \{0,1\} & \textit{(trivial equilibrium)} \end{cases} \tag{16}$$

Theorem 8 indicates that the conditions for internal equilibrium are described by the *zeros* of the function $W_1(\phi) - W_0(\phi)$, as depicted in Fig. 2, and, by the threshold equation, Eq. (6), ϕ has dynamical dependence only on $\overline{s}$. It follows that only certain values of $\overline{s}$ support internal equilibrium, and each value corresponds to a hyperplane in state space (Remark 4).

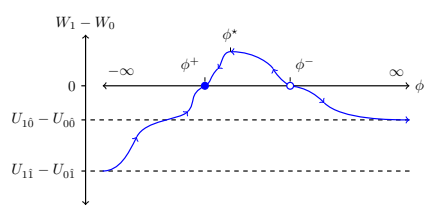

Figure 2: $W_1(\phi) - W_0(\phi)$ (blue curve) is a strictly quasi-convex (concave) function of ϕ. $\phi^\star$ denotes the unique local extremum. The direction of the arrows is a consequence of Theorem 7 and Corollary 1.1.

Theorem 9. *$W_1(\phi) - W_0(\phi)$ is strictly quasi-concave in ϕ. This guarantees that no more than two zeros of the function $W_1 - W_0$ exist.*

We denote the possible zeros of $W_1 - W_0$ as ϕ^+ and ϕ^-, where the sign in the superscript indicates the local slope of the function. These zeroes correspond to parallel hyperplanes in state space that comprise all interior equilibria of the system. Whether $\phi^\pm$ corresponds to an (un)stable equilibrium hyperplane may be determined by Theorem 7 and the sign of $\frac{\partial}{\partial \phi}(W_1 - W_0)$: Only $\overline{s}(\phi^+)$ is stable, and we will verify this fact with linear stability analysis.

Theorem 10. *If the state of the system asymptotically approaches an internal equilibrium, the nullity of total qualification disparity $\|D[t]\|_p$, where $p \geq 1$, is preserved in the limit of infinite time.*

$$\lim_{t' \to \infty} (W_1 - W_0) = 0 \implies \Big(\big\|D[t]\big\|_p = 0 \iff \lim_{t' \to \infty} \big\|D[t']\big\|_p = 0\Big) \tag{17}$$

Theorem 10 formalizes the critical observation that any state that attracts to the stable equilibrium hyperplane, unless initially free from qualification disparity, will forever exhibit some total qualification disparity. This is a more robust notion of the **persistence of disparity** in our system than Theorem 7.

3.2 Stability

For our regional banking example, we may imagine that qualification rates settle into a stable pattern in which some communities have a higher average qualification rate than others. How robust is this pattern of inequality to small fluctuations of qualification rates? Using **linear stability analysis** (*i.e.*, linearizing the response of the system to small perturbations about equilibrium and asking "do perturbations amplify or dissipate?"), we show that only the ϕ^+-hyperplane acts as a stable attractor.

First, let us denote the evaluation of an expression *at equilibrium* by placing a vertical line to the right of the expression with "eq" as a subscript. In light of Theorem 8 and Eq. (7), we also introduce the shorthand W_{eq} to denote an equilibrium value of W_1, W_0, or, equivalently, any $\overline{W}_g$. It should be noted that the value of $W_{\text{eq}} \in [0, \infty)$ still depends on the particular equilibrium state of the system.

$$W_{\text{eq}} := W_0\Big|_{\text{eq}} = W_1\Big|_{\text{eq}} = \overline{W}_g\Big|_{\text{eq}} \quad \forall g \in \mathcal{G} \tag{18}$$

We linearize the system at equilibrium by constructing the Jacobian $J \in \mathbf{R}^{n \times n}$ corresponding to discrete time-evolution and identifying its eigenvectors and eigenvalues:

$$J := \left[\frac{\partial \mathbf{r}}{\partial \delta(1,2)} \; \frac{\partial \mathbf{r}}{\partial \delta(2,3)} \; \cdots \; \frac{\partial \mathbf{r}}{\partial \delta(n-1,n)} \; \frac{\partial \mathbf{r}}{\partial \overline{s}} \right] \tag{19}$$

where $\mathbf{r}$, the state vector in $(D, \overline{s})$ coordinates, is interpreted as a column vector.

Theorem 11. *The Jacobian J simplifies to a scalar multiplied by a matrix with a single non-zero column $\mathbf{v}$ in the last position.*

$$J\Big|_{\text{eq}} = \frac{1}{W_{\text{eq}}} \left(\frac{d\phi}{d\overline{s}} \right) \left(\frac{d}{d\phi}(W_1 - W_0) \right) \left[\mathbf{0}^{(n \times n-1)} \middle| \mathbf{v} \right], \quad \mathbf{v} := \begin{bmatrix} \delta(1,2)(1-s_1-s_2) \\ \delta(2,3)(1-s_2-s_3) \\ \cdots \\ \delta(n-1,n)(1-s_{n-1}-s_n) \\ \sum_{g \in \mathcal{G}} \mu_g s_g (1-s_g) \end{bmatrix} \tag{20}$$

The eigenvalues of J plays an important role in determining the stability of the system:

Corollary 11.1. *At equilibrium: any state vector $\mathbf{r}$ with zero $\overline{s}$ component is an eigenvector of J with eigenvalue 0, and $\mathbf{v}$ is an eigenvector of J with eigenvalue λ, given by*

$$\lambda := \left(\sum_{g \in \mathcal{G}} \mu_g s_g (1 - s_g) \right) \frac{1}{W_{\text{eq}}} \left(\frac{d\phi}{d\overline{s}} \right) \left(\frac{d}{d\phi}(W_1 - W_0) \right) \Bigg|_{\text{eq}} \tag{21}$$

Perturbing a state vector $\mathbf{r}$ at internal equilibrium by altering any combination of coordinates specified by D while leaving $\overline{s}$ fixed corresponds to motion along an equilibrium hyperplane; such motion may occur in neutral equilibrium as a result of the null eigenvalues of the perturbing vector, for which the $\overline{s}$ component is zero. An equilibrium state is stable to perturbations normal to the equilibrium hyperplane iff λ is negative (and, in discrete-time, > -2, to prevent repeated over-corrections) [41]:

Corollary 11.2. *As a consequence of Corollary 1.1, which states $\frac{d\phi}{d\overline{s}} < 0$, the eigenvalue λ in Eq.* (21) *is negative, (and the associated equilibrium hyperplane stable) iff $\frac{d}{d\phi}(W_1 - W_0)|_{\text{eq}} > 0$. This prescribes precisely the value ϕ^+ for the stable equilibrium hyperplane.*

4 Interventions

We explore "fairness interventions" for the dynamical system we have characterized. We first observe that our setting so far, requiring a group-independent feature threshold ϕ and distributions q_y, automatically satisfies one commonly cited standard of normative present fairness:

Definition 12. Equalized Odds [12–14] requires that a classifier's decisions $\hat{Y}$, given by policy π, misclassify (un)qualified agents at equal rates across groups:

$$\forall g, h \in \mathcal{G}. \, \forall y, \hat{y} \in \{0, 1\}. \quad \Pr(\hat{Y} = \hat{y} \mid Y = y, G = g) = \Pr(\hat{Y} = \hat{y} \mid Y = y, G = h) \tag{22}$$

Theorem 13. *By direct consequence of Assumption 2, which guarantees the group-independence of q_y, for group-specific threshold policies, the equivalence of feature thresholds is necessary and sufficient for Equalized Odds.*

Theorem 13 indicates that Equalized Odds (*e.g.*, a bank rejects (accepts) (un)qualified loan applicants at group-independent rates) is completely ineffectual for long-term fairness in our model, since we have already shown how disparities persist under a policy satisfying its criteria. We next ask whether an intervention that perturbs the group-independent threshold ϕ near the ϕ^+-hyperplane, which we interpret as a **universal subsidy** (or penalty), can reform qualification disparities.

Theorem 14. *To first-order approximation, perturbation of a group-independent ϕ at equilibrium induces motion parallel to the eigenvector* $\mathbf{v}$.

As a consequence of Theorem 14, while $\mathbf{v}$ need not be orthogonal to the equilibrium hyperplane, and a universal subsidy may improve qualification disparity while applied (settling on a new equilibrium hyperplane with different, though persistent disparities), the system is stable to such perturbations as characterized by linear system response and will relax to the original equilibrium state when the intervention is removed. To permanently change qualification disparities, such an intervention, temporarily applied, must rely on the **non-linear response** of the system and is therefore liable to require *large* perturbations to the classifier's threshold ϕ. This finding compels us to consider interventions with group-dependent threshold perturbations—or group-dependent thresholds. To this end, we hereafter generalize our classifier such that it independently classifies each group g according to a group-specific threshold ϕ_g. We denote the vector of these thresholds as $\Phi \coloneqq (\phi_1, \phi_2, ..., \phi_n)$ and assume that, prior to some perturbative intervention, $\phi_g = \phi$ for each $g \in \mathcal{G}$.

Definition 15. Demographic Parity [10, 11] requires that a classifier's decisions $\hat{Y}$, given by policy π, are positive ($\hat{Y} = 1$, *e.g.*, accepting a loan application) at equal rates for all groups:

$$\forall g, h \in \mathcal{G}. \quad \Pr(\hat{Y} = 1 \mid G = g) = \Pr(\hat{Y} = 1 \mid G = h) \tag{23}$$

Theorem 16. *Demographic parity requires sign-heterogeneous, group-dependent perturbations to an initial, group-independent threshold ϕ when $\pi(\phi)$ is non-trivial (does not uniformly accept (reject)).*

Unfortunately, satisfying demographic parity is a **non-convex** optimization problem [42] and in our case requires the solution of a differential equation dependent on q_y (see Appendix). We therefore rely on numerical simulation, rather than analytical tools, to evaluate this intervention for our system.

Feedback control Arbitrary state transitions in the equilibrium hyperplane may be permanently effected by group-dependent perturbations to Φ, which we derive from linear system response at equilibrium. Specifically, to diminish a specific qualification distance $\delta(g, g+1)$ for given g, Φ may be perturbed by a vector quantity $\Delta_g \Phi = (\Delta_g \phi_1, \Delta_g \phi_2, ..., \Delta_g \phi_n)$.

Theorem 17. *On the stable internal equilibrium hyperplane, infinitesimal perturbation of Φ by*

$$\Delta_g \Phi \coloneqq -\epsilon \delta(g, g+1) \Big(\frac{\alpha_g}{s_1(1-s_1)}, ..., \frac{\alpha_g}{s_g(1-s_g)}, \frac{\beta_g}{s_{g+1}(1-s_{g+1})}, ..., \frac{\beta_g}{s_n(1-s_n)} \Big) \tag{24a}$$

$$\alpha_g \coloneqq (\mu_{g+1} + \mu_{g+2} + ... + \mu_n), \qquad \beta_g \coloneqq -(\mu_1 + \mu_2 + ... + \mu_g) \tag{24b}$$

*will induce motion in the system preserving $\overline{s}$ and each $\delta(h, h+1)$ for $h \neq g$. The value of $\delta(g, g+1)$ will be diminished by a ratio proportional to the **strength parameter** $\epsilon > 0$.*

Perturbations of the form $\Delta_g \Phi$ may be composed linearly for multiple values of g. In particular, when ϵ is a universal quantity, we may determine the total perturbation to Φ necessary to simultaneously and proportionately decrease all qualification distances for any given state on the stable equilibrium hyperplane. Let us denote this total perturbation as $\Delta\Phi \coloneqq \sum_{g \in \mathcal{G}} \Delta_g \Phi = (\Delta\phi_1, \Delta\phi_2, ..., \Delta\phi_n)$. Component-wise, $\Delta\Phi$ is given by

$$\Delta\phi_g = \frac{-\epsilon}{s_g(1-s_g)} \left(\sum_{h=g}^{n-1} \alpha_h \delta(h, h+1) + \sum_{h=1}^{g-1} \beta_h \delta(h, h+1) \right) \tag{25}$$

We note that the proposed feedback control mechanism depends only on the known constants μ_g and feedback in terms of current qualification distances δ. In addition, the force of the intervention can be tuned by setting the strength parameter ϵ. Finally, we remark that this mechanism can be composed with global perturbations of ϕ, in the manner of Theorem 14, to intervene without rejecting any agents that would have been accepted under group-independent policies.

We compare interventions by appeal to simulation, choosing a setting that guarantees a single, stable average qualification rate $\overline{s}^\star$ under group-independent policies (**GI**) (Fig. 3). We consider trade-offs

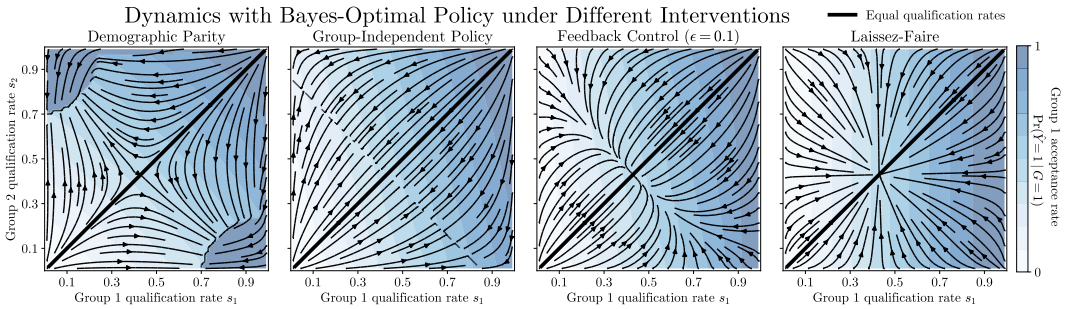

Figure 3: Simulated dynamics for two groups of equal size, subject to different global interventions. Streamlines approximate system time evolution. q_0 and q_1 are Gaussians with unit variance and have means -1 and 1, respectively. Other examples and rendered dynamical variables are provided in the Appendix. For this example, $(U_{0\hat{0}} = 0.1; U_{0\hat{1}} = 5.5; U_{1\hat{0}} = 0.5; U_{1\hat{1}} = 1.0; V_{0\hat{0}} = 0.5; V_{0\hat{1}} = -0.5; V_{1\hat{0}} = -0.25; V_{1\hat{1}} = 1.0)$.

between *normative present fairness* ($\mathbf{F_{NP}}$) (*e.g.*, demographic parity (**DP**) or equalized odds (**EO**)) and *long-term fairness* ($\mathbf{F_{LT}}$), for which the dynamics must converge to the line demarcating equal qualification rates. Darker shading (blue) implies a higher absolute acceptance rate for Group 1, which, by the setting's symmetry, is the same for Group 2 when reflected across the aforementioned line; reflectional asymmetry violates DP. Under compulsory DP (first pane), the system violates $\mathrm{F_{NP}}$, settling into a "patronizing equilibrium under affirmative action", as coined by Coate and Loury [21], in which agents from a less-qualified group are *patronized* (*e.g.*, granted loans despite nonqualification) by the classifier (*e.g.*, the dark blue upper-left corner with low qualification and high acceptance rates for Group 1). States under GI (second pane), which satisfies the EO notion of $\mathrm{F_{NP}}$ (*cf.* Theorem 13), converge to a line (hyperplane, *cf.* Remark 4) while preserving qualification disparities (*cf.* Theorem 10). $\mathrm{F_{LT}}$ is expected from a **laissez-faire** (**LZ**) policy (last pane), which adopts group-specific policies and thus decouples all group dynamics: Each group must converge to $\overline{s}^\star$ separately. Still, LZ satisfies neither the DP (by reflectional asymmetry) nor EO (by Theorem 13) notions of $\mathrm{F_{NP}}$. In contrast, feedback control (last pane) achieves $\mathrm{F_{LT}}$ at the expense of small, parametric violations of $\mathrm{F_{NP}}$ (EO) with small ϵ (See Appendix for plots of classifier error rates in this setting).

5 Discussion and limitations

The novelty of our contribution is the demonstration of persistent qualification rate disparities in an equitable setting consistent with plausible mechanisms of population response—caused by careless deployment of machine learning and myopic fairness interventions. We submit that, given the many charitable assumptions of our model to achieve perfect structural equality between groups, any reasonable fairness intervention should succeed in responsibly rectifying disparities here, if anywhere. Moreover, we have laid bare inherent tensions that can exist between the *means* and *ends* of fairness considerations in a dynamical context, demonstrating potential incompatibility of immediate and long-term notions of fairness.

We acknowledge that our model is simplistic, but such simple cases must be well-understood as a first step towards more equitable models of population response. We regard strictly segregated groups as the most tenuous assumption of our model and conjecture that even relatively weak inter-group exchange of strategies should lead to long-term fairness in our default setting. Nonetheless, we believe that a program based on incomplete agent information can successfully endogenize persistent disparities more robustly. Specifically, future work may consider multiple classifiers with different task domains affecting a common population; we expect this to readily endogenize broken symmetries between group environments and conditions. We also trust that voluntary participation, as considered by Zhang et al. [31], may be modelled as an additional strategy within our framework. Regarding empirical falsifiability, we note that the dynamics of social disparity are not exclusive to algorithmic classifiers [27], and ask whether our model's predictions may be contrasted with existing and historical resource allocation problems.

Finally, we invite readers to consider both our model and explicit application of control theory to society through algorithmic classification with care, as our work is intended to *reform* the misapplication of machine learning, inequitable modelling, and myopic notions of fairness—not perpetuate them.

References

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

[2] Karen Hao. The coming war on the hidden algorithms that trap people in poverty. *MIT Technology Review*, 2020.

[3] Cade Metz and Adam Satariano. An Algorithm That Grants Freedom, or Takes It Away. *The New York Times*, 2020.

[4] Derek Newton. From admissions to teaching to grading, AI is infiltrating higher education. *The Hechinger Report*, 2021.

[5] Joe Hernandez. A Military Drone With A Mind Of Its Own Was Used In Combat, U.N. Says. *National Public Radio*, 2021.

[6] Kate Crawford and Ryan Calo. There is a blind spot in ai research. *Nature News*, 538(7625):311, 2016.

[7] Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 224–232. ACM, 2018.

[8] Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. Predictably unequal? the effects of machine learning on credit markets. *The Effects of Machine Learning on Credit Markets*, 2018.

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

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

[11] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

[12] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning, 2016.

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

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

[15] 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.

[16] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[17] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 228–238. Curran Associates Inc., 2017. ISBN 9781510860964.

[18] Berk Ustun, Yang Liu, and David Parkes. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, pages 6373–6382. PMLR, 2019.

[19] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017.

[20] Maximilian Kasy and Rediet Abebe. Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 576–586, 2021.

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

[22] 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.

[23] Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellström, Kun Zhang, and Cheng Zhang. How do fair decisions fare in long-term qualification? *arXiv preprint arXiv:2010.11300*, 2020.

[24] 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. *the International Conference on Machine Learning (ICML)*, 2019.

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

[26] 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.

[27] 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 on World Wide Web*, pages 1389–1398. International World Wide Web Conferences Steering Committee, 2018.

[28] Hussein Mouzannar, Mesrob I Ohannessian, and Nathan Srebro. From fair decision making to social equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 359–368. ACM, 2019.

[29] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018.

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

[31] Xueru Zhang, Mohammad Mahdi Khalili, Cem Tekin, and Mingyan Liu. Group retention when using machine learning in sequential decision making: the interplay between user dynamics and fairness. In *Advances in Neural Information Processing Systems*, pages 15243–15252, 2019.

[32] Wei Tang, Chien-Ju Ho, and Yang Liu. Bandit learning with delayed impact of actions. *arXiv preprint arXiv:2002.10316*, 2020.

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

[34] Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C Parkes. Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875*, 2017.

[35] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1617–1626. JMLR. org, 2017.

[36] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *Conference of Fairness, Accountability, and Transparency*, 2018.

[37] Jonas Björnerstedt and Jörgen W Weibull. Nash equilibrium and evolution by imitation. Technical report, IUI Working Paper, 1994.

[38] Peter D Taylor and Leo B Jonker. Evolutionary stable strategies and game dynamics. *Mathematical biosciences*, 40(1-2):145–156, 1978.

[39] Richard Dawkins. *The selfish gene*. 1976.

[40] Daniel Friedman and Barry Sinervo. *Evolutionary games in natural, social, and virtual worlds*. Oxford University Press, 2016.

[41] Steven H Strogatz. *Nonlinear dynamics and chaos with student solutions manual: With applications to physics, biology, chemistry, and engineering*. CRC press, 2018.

[42] Yongkai Wu, Lu Zhang, and Xintao Wu. On convexity and bounds of fairness-aware classification. In *The World Wide Web Conference*, pages 3356–3362, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] In the supplementary material
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] In the attached source code "readme" file
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]