Cultural Evolution by Unconscious Selection: Persistent Qualification Disparities and Interventions

Anonymous Author(s)
Affiliation
Address
email

Abstract

Appropriately modeling the dynamics of group-level disparities in machine learning without assuming a static, structurally biased setting remains an open problem: one that has induced previous research on fairness interventions to contend with symptomatic treatment of underlying, inoperable social issues. In this paper, we appeal to the replicator equation, an established model for evolutionary phenomena without mutation, to model how the population density of qualification, i.e., some elective binary status, evolves across isolated subpopulations. We assume that the evolutionary fitness of (non)qualification is influenced by the predictions of an optimal classifier repeatedly retrained on the global population. While assuming that agent behaviors, classifier utilities, and label-conditioned feature distributions are group-independent, we identify a set of non-trivial equilibrium states at which differences in qualification rates between subpopulations can persist indefinitely. We next compare the effects of commonly proposed fairness interventions on this dynamical system to a new feedback control mechanism capable of permanently eliminating group-level qualification disparities. We conclude by discussing the limitations of our model and these findings and by outlining future work.

1 Introduction

The emergent use of automated classifiers for socially consequential decisions has raised both technical challenges and ethical concerns, particularly regarding the dynamics of systemic inequality [7, 3, 2, 4, 14]. While prior literature in algorithmic fairness has focused on mitigating statistical disparities between subpopulations with *fairness interventions* such as statistical parity [5, 9], equal opportunity [16], or envy-freeness [30, 10, 28], such work typically assumes a static setting to evaluate these techniques. Recent work has added time-dependence by modeling changes to underlying qualification in response to machine classification but has typically considered unstable dynamics [22, 11, 25] inconsistent with persistent, real-world disparities, or has assumed a fixed, structurally unequal setting that does not respond to changing conditions [17, 34].

Our primary contribution is a general dynamical model of group-dependent label transitions consistent with real-world, persistent inequalities, free from structurally unfair assumptions. We formulate this model in Section 2. In particular, we assume that group membership does not affect agent behaviors ([22, 34]), label-conditioned feature distributions ([34]), nor classifier utilities. In our formulation, groups vary only in relative size and initial qualification rates. We model updates to agent qualification by appeal to replicator dynamics Section 2.2, which capture the dynamics of evolution without mutation [13]. At the same time, we assume the classifier regularly updates its policy to maximize its immediate utility with perfect knowledge of the global joint distribution of features and labels. The mutual recurrence of these updates to the population and the classifier's

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

policy defines our dynamical system. The classifier in this model unconsciously influences the evolutionary success of qualification as a strategy in each group.

Our second contribution is a rigorous examination of this dynamical system, including a complete characterization of its equilibrium states with linear stability analysis (Section 3). 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 12). In this sense, we claim that qualification disparity persists indefinitely.

Our final contribution is a study of fairness interventions as applied to our model. We consider a global perturbation to the classifier's policy, equal opportunity [16], demographic parity [31], and a new mechanism based on feedback control (Section 4). We show that the feedback control intervention differs substantially from previously proposed, group-dependent fairness interventions and can eliminate qualification disparity from the system. We conclude by discussing the limitations of our model and these findings and by outlining future work.

1.1 Related work

Our work chiefly contributes to the literature on fairness in machine learning. Established work in this field has proposed several intervention techniques to guarantee (variously defined and sometimes mutually incompatible) *fair* outcomes from machine learning classifiers [5, 16, 10, 1]. For example, [31] introduced corrections to a classifier's loss function to minimize the mutual information between group membership and classification rates, thereby enforcing demographic parity, an intervention we evaluate with our model in Section 4. Later studies have demonstrated the (im)possibility of achieving fairness in dynamic learning settings, *e.g.*, in a Multi-Armed Bandit setting [20, 19, 15, 23].

The most relevant publications to our current contribution are those which have highlighted the importance of studying the dynamics and long-term consequences for fairness in machine learning. For instance, [22] revealed that deploying a fairness-aware classifier can have unintended consequences through delayed impact, while [12] studied the positive feedback effects of bias in predictive policing. [34] modeled updates to agent qualification rates in response to classification using a Markov transition rather than replicator dynamics, but still assumed fixed, structural inequalities.

Relevant literature on replicator dynamics includes the original conception of the model in a game-theoretic context, inspired by a characterization of evolutionarily stable strategies [26]. Recent work has appealed to this model to predict the equilibrium states for multiagent systems utilizing individual reinforcement learning algorithms [27].

Note that we defer all proof to the supplementary material.

2 Formulation

We first define the properties of our primary objects: *agents* (individuals), *groups* (subpopulations), and a *classifier*, where we have chosen nomenclature for these objects to mirror past literature, before elaborating on the modeled behavior of each.

Agents We consider a population of (countably many) **agents** distributed between $n \geq 1$ groups $\mathcal{G} = \{1, 2, ..., n\}$ with known frequencies $\mu_g \in (0, 1)$. We reserve the symbols g, h, i, j as indices corresponding to elements of $\mathcal{G}$. Letting the random variable G represent the unknown group membership of a randomly selected agent, we define

$$\mu_g \stackrel{\text{def}}{=} \Pr(G = g), \quad \sum_{g \in \mathcal{G}} \mu_g = 1, \quad \boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_n) \tag{1}$$

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

In addition to group membership G, each agent is characterized by a binary label $Y = \{0, 1\}$, where we describe an agent with label value $Y = 1$ as *qualified* (for some semantically *positive* classification), and a real-valued feature $X \in (-\infty, \infty)$. We reserve the symbols y, z as indices corresponding to the Boolean elements of Y.

Assumption 1. *We assume that the feature X depends on the label Y according to a probability density function q that is independent of group membership G.*

$$q_y(x) \overset{def}{=} p_X(x \mid Y = y), \quad y \in \{0, 1\} \tag{2}$$

By consciously avoiding structural inequality in our assumptions, our observation of persistent disparities in this model becomes significant.

Assumption 2. *First, we assume that q is differentiable and strictly positive: $\forall x, y. \quad q_y(x) \in (0, \infty)$. With a loss of generality that may be constrained to vanishingly small perturbations of q, we further choose to (re)order the values of X such that $q_1(x)/q_0(x)$ is strictly increasing in X.*

$$\frac{d}{dx}\left(\frac{q_1(x)}{q_0(x)}\right) > 0 \tag{3}$$

Groups We carefully consider the interpretation of a **group** of agents in Section 2.2, but for now, we may characterize a group g by the fraction of agents in g that are qualified; we refer to this fraction as the group's *qualification rate* s_g. We denote the **state** variable of the system as $\mathbf{s}$ and denote the global qualification rate as $\overline{s}$.

$$s_g \overset{\text{def}}{=} \Pr(Y = 1 \mid G = g), \quad \mathbf{s} = (s_1, s_2, ..., s_n), \quad \overline{s} \overset{\text{def}}{=} \sum_{g \in \mathcal{G}} \mu_g s_g = \langle \boldsymbol{\mu}, \mathbf{s} \rangle \tag{4}$$

Assumption 3. *For all groups, $s_g \in (0, 1)$;* i.e.*, no group is completely (un)qualified.*

It follows from $\mu_g \in (0, 1)$ that, as a convex combination of each s_g, $\overline{s} \in (0, 1)$.

The classifier We lastly introduce a **classifier** which observes the feature X of each agent and is tasked with predicting the agent's correct label Y. The classifier achieves this by generating a deterministic [1] policy π that maps each feature value X to a binary prediction value $\hat{Y} \in \{0, 1\}$. A prediction of $\hat{Y} = 1$ corresponds to *accepting* an agent, while $\hat{Y} = 0$ implies *rejecting* the same.

Assumption 4. *We assume that the classifier knows the true distribution $\Pr(Y \mid X)$ for the entire population.*[2]

Assumption 5. *We assume that the classifier expresses risk-neutral preferences in selecting a policy π that maximizes the expected value of a utility function u. We also assume that u is linear in the outcome fractions $\Pr(Y = y, \hat{Y} = z)$ resulting from the policy π, with coefficients defining a matrix $V \in \mathbf{R}^{2\times 2}_{\geq 0}$. Finally, we assume that the utility to the classifier for (in)correct predictions is independent of the feature value X and group membership G.*

$$\hat{Y} \overset{def}{=} \pi(X), \quad u(\pi) \overset{def}{=} \sum_{y,z=0}^{1} V_{yz} \Pr_{\hat{Y} \sim \pi(X)}(Y = y, \hat{Y} = z) \tag{5}$$

Theorem 1. *Discounting sets of measure zero, the u-maximizing policy π is parameterized by a single, **probability threshold** $\theta \in [0, 1]$ such that $\pi(x) = \begin{cases} 1 & \Pr(Y = 1 \mid X = x) > \theta \\ 0 & \textit{otherwise} \end{cases}$ where $\theta = \frac{V_{00} - V_{01}}{V_{11} - V_{11} + V_{00} - V_{01}}$.*

Assumption 6. *We assume that the optimal policy is not universally trivial. That is $\theta \in (0, 1)$.*

Theorem 2. *The u-maximizing policy π is a threshold classifier, where the classifier's **feature threshold** $\phi \in [-\infty, \infty]$ is defined such that $\pi(x) = \begin{cases} 1 & x > \phi \\ 0 & \textit{otherwise} \end{cases}$. Furthermore, ϕ depends only on the global qualification rate $\overline{s}$ as*

$$\frac{q_1(\phi)}{q_0(\phi)} = \left(\frac{\theta}{1 - \theta}\right)\left(\frac{1 - \overline{s}}{\overline{s}}\right) \tag{6}$$

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

[1] Our framework may treat a non-deterministic classifier by subsuming all non-determinism into the probability density function q_y, so long as Assumption 2 remains satisfied.

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

We conclude our introduction of the classifier by observing that it becomes more discerning (ϕ increases) when fewer individuals qualify ($\overline{s}$ decreases):

Corollary 2.1. *The classifier's feature threshold ϕ responds inversely to $\overline{s}$:* $\frac{d\phi}{d\overline{s}} < 0, \quad \frac{d\overline{s}}{d\phi} < 0.$

Corollary 2.1 will be useful in proving Corollary 13.2, which determines which subset of equilibrium states constitute an attractor.

2.1 Time-dependence

Assumption 7. *μ_g, q_y, and V are time-independent for all $y \in \{0, 1\}, g \in \mathcal{G}$. Furthermore, prior assumptions regarding the classifier's policy generation process and knowledge of* $\Pr(Y \mid X)$ *also hold independently for each time step.*

Thus far, we have established relationships that define the same Markov chain for any randomly selected agent, reflecting conditional independence conditions from Assumption 1 and Assumption 5.

$$G \overset{s_g}{\to} Y \overset{q}{\to} X \overset{\pi}{\to} \hat{Y} \tag{7}$$

We next introduce time-dependence through a feedback mechanism known as *label shift* [32, 21], whereby the qualification rate s_g in each group changes in response to the classifier policy π.

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

$$\begin{tikzcd} \mathbf{s}[t] \arrow[r, "\theta, q"] \arrow[rrr, bend right] & \pi(\phi[t]) \arrow[r, "U, Q"] & (W_0[t], W_1[t]) \arrow[r] & \mathbf{s}[t+1] \end{tikzcd}$$

Figure 1: Functional dependence of dynamical variables

2.2 Replicator dynamics

When modeling changes to qualification rates or accounting for observed disparities in reality, it is common to appeal to individual costs and utility: For example, we might judge that, due to a lack of opportunities or resources, members of some disadvantaged group (for which s_g is relatively low) incur a higher cost in attaining qualification and are therefore less likely to (choose to) become qualified [22]. Alternatively, we might speculate that certain groups are inherently more or less capable or motivated to secure qualification [34]. Either type of assumption may enter a utility-based model via the same parametric inputs, but either may be contentious.

Moreover, when relying on explanations of personal utility, an explicit mechanism by which systemic disparity *directs* divergence in personal utilities is still required to model the *long-term* dynamics of the system. Without a way to revise or update the "inherent" differences between groups as time progresses, nor the judgment to assume permanent, structurally unequal parameter values between groups, our approach is to remain agnostic about individual motivations and costs and focus instead on the (evolutionary) fitness W of (non)qualification as a behavioral strategy within a community.

Our approach acknowledges and embraces an understanding that machine learning classifiers can promote certain behaviors in a population *unconsciously*, that is, without intentionally selecting for them, in much the same way that early plant domestication may have been driven chiefly by the selective pressures of a human-managed environment rather than conscious trait selection [35].

As commonly encountered in the literature of evolutionary game theory, the **fitness** $W_y^g \in [0, \infty)$, (where g is used as a superscript rather than an exponent), of strategy $Y = y$ within any closed population g may be defined up to a universal, positive constant of proportionality such that the (discrete) **replicator equation** governs the time-evolution of the frequency of the strategy in g [13].

$$s_g[t+1] = s_g[t]\frac{W_1[t]}{\overline{W}_g[t]}, \quad (1 - s_g[t+1]) = (1 - s_g[t])\frac{W_0[t]}{\overline{W}_g[t]}, \quad \overline{W}_g \overset{\text{def}}{=} W_1 s_g + W_0(1 - s_g) \tag{8}$$

[3] The generalization of the replicator equation to continuous time is provided by [13]

To motivate intuitive interpretation of these equations, a model of genetic reproduction permits substitution of the average number of offspring of genotype y in population g for W_y^g [13]. The time-independence of μ_g forces us to interpret the average fitness in group g, $\overline{W}_g$, as a group-dependent normalization constant for W_1 and W_0 rather than the rate of growth for population g. Finally, an equivalent, though less standard definition suggests the *ratio* of any two fitness values W_y^g as the primary objects of consideration:

$$\frac{s_g[t+1]}{(1-s_g[t+1])} = \frac{W_1^g[t]}{W_0^g[t]} \frac{s_g[t]}{(1-s_g[t])} \tag{9}$$

It is significant that such dynamics apply to *closed* populations, as this both informs and restricts our interpretation of groups: Our model requires an understanding of the boundaries between groups as functionally impermeable to the exchange of qualification strategies:

Assumption 8. *We assume that each group g has the properties of an isolated population in which* qualification*, as a strategy and/or meme*[4]*, competes with non-qualification free from exchange with other groups.*

We contrast this assumption with the observation that groups in prior literature are frequently defined with respect to "sensitive attributes" such as race, religion, sex, *etc.* [22, 33, 34, 5, 16]. Such attributes can be dubious proxies for meaningful divisions and machine learning based on such group designations can cause active harm [6]. By modeling groups as isolated populations, we do *not* project such divisions onto these designations; rather we *define* what we mean by a group by the extent to which it satisfies Assumption 8.

Our following assumptions ensure that, despite the divisions between groups, the fitness of either qualification strategy remains independent of group membership while responding to the classifier's policy:

Assumption 9. *The fitness W_y^g of strategy $Y = y$ in group g is the average fitness of individuals in that group who select that strategy, where we further assume a time-, feature- and group-independent average fitness U_{yz} for agents who select strategy $Y = y$ and are classified as $\hat{Y} = z$.*

$$W_y^g \stackrel{def}{=} \sum_{z=0}^{1} \Pr(\hat{Y} = z \mid Y = y, G = g) U_{yz} \tag{10}$$

So as to continue avoiding degenerate cases, let us assume that:

Assumption 10. *$U_{01} \neq U_{00}$ and $U_{11} > U_{10}$. That is, the fitness of either label is sensitive to classification, and, for qualified individuals, positive classification increases fitness relative to rejection.*

Definition 3. As a consequence of Theorem 2, we define the **cumulative distribution functions**

$$Q_y(\phi) \stackrel{\text{def}}{=} \int_{-\infty}^{\phi} q_y(x)\, dx = \Pr(\hat{Y} = 0 \mid Y = y), \quad y \in \{0, 1\} \tag{11}$$

Theorem 4. *The fitness W_y^g of strategy $Y = y$ in each group g is group-independent. We will hereafter denote this shared label-fitness as W_y (without superscript).*

$$\forall g \in \mathcal{G}. \quad W_y^g = W_y = U_{y1} + (U_{y0} - U_{y1}) Q_y(\phi), \quad y \in \{0, 1\} \tag{12}$$

3 Dynamics

We now consider how the dynamical system we have defined evolves in time by coupling the replicator equation (Eq. (8)), which allows us to calculate the next value of $\overline{s}$ as a function of ϕ, and the threshold equation (Eq. (6)), which yields ϕ as a function of $\overline{s}$. We begin by creating a useful set of coordinates to compliment $\overline{s}$ and track qualification disparities. We then note the importance of $W_1(\phi) - W_0(\phi)$ to the overall dynamics of the system, and use it to identify all equilibrium states.

[4] in the sense originally coined by Dawkins [8]

Definition 5. Define the (signed) **qualification distance** from group h to group g as

$$\delta(g,h) \stackrel{\text{def}}{=} s_g - s_h, \quad g,h \in \{1,2,...,n\} \tag{13}$$

To describe the equilibrium states of the system simply, we first perform a linear, non-orthogonal change of coordinates from the vector of group-specific qualification frequencies $\mathbf{s} = (s_1, s_2, ..., s_n)$ to a vector that comprises $\overline{s}$ and as set D of $(n-1)$, linearly-independent qualification distances between sequential pairs of subpopulations:

$$D \stackrel{\text{def}}{=} \Big\{\delta(1,2),\ \delta(2,3),\ ...,\ \delta(n-1,n)\Big\} \tag{14}$$

D and $\overline{s}$ together yield a complete set of coordinates to describe the state of the dynamical system, which we may exchange for the original qualification rates via linear operations:

$$s_g = \overline{s} + \sum_{h=g}^{n-1} \delta(h,h+1) - \sum_{h=1}^{n-1}\sum_{k=1}^{h} \mu_k \delta(h,h+1) \quad \forall g \in \mathcal{G} \tag{15}$$

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

Remark 6. By the linear definition of $\overline{s}$ in Eq. (4), we may describe all internal equilibrium states with a specific value of $\overline{s}$ as a **hyperplane**.

Remark 7. The nullity of $\|D[t]\|_p, p \geq 1$ is preserved in time.

$$\Big(\sum_{g=1}^{n-1} \big|\delta(g,g+1)[t]\big|\Big)^{1/p} = 0 \iff \Big(\sum_{g=1}^{n-1} \big|\delta(g,g+1)[t+1]\big|\Big)^{1/p} \tag{16}$$

Remark 7 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. Note that this statement is insufficient to address the limit $t \to \infty$, however. For a stronger result that includes this limit, we characterize the system's equilibrium states by first noticing the importance of $W_1(\phi) - W_0(\phi)$ in determining the dynamics of the system:

Theorem 8. *Disregarding boundary states by Assumption 3,*

$$\Big(\overline{s}[t+1] - \overline{s}[t]\Big) > 0 \iff (W_1(\phi) - W_0(\phi)) > 0 \tag{17}$$

This observation leads directly into a characterization of equilibrium.

3.1 Equilibrium

Definition 9. The system as a whole is **at equilibrium** when, for all $g \in \mathcal{G}$ simultaneously, s_g is stationary in time:

$$\text{at equilibrium} \stackrel{\text{def}}{\iff} \forall g \in \mathcal{G}.\ \exists t_0 \text{ s.t. } \forall t \geq t_0, \quad s_g[t] = s_g[t_0] \tag{18}$$

Theorem 10. *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.*

$$\textit{at equilibrium} \iff \begin{cases} W_1 = W_0 & \textit{(internal equilibrium)} \\ \forall g \in \mathcal{G}. \quad s_g \in \{0,1\} & \textit{(trivial equilibrium)} \end{cases} \tag{19}$$

We may describe the conditions for internal equilibrium by the *zeros* of the function $W_1(\phi) - W_0(\phi)$, as depicted in Fig. 2.

Theorem 11. *$W_1(\phi) - W_0(\phi)$ is strictly quasi-convex in ϕ if $U_{00} > U_{01}$ and strictly quasi-concave if $U_{00} < U_{01}$. This guarantees that not 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 corresponds to the local slope of the function. Since ϕ depends on the state of the system only via $\overline{s}$, according to Eq. (6), it follows that only specific values of $\overline{s}$ permit equilibrium, each corresponding to a hyperplane (Remark 6) in state space. As we will see by performing linear stability analysis at equilibrium, whether a specific value of $\overline{s}$ corresponds to a stable or unstable equilibrium hyperplane is determined by the sign of $\frac{\partial}{\partial \phi}(W_1 - W_0)$, and only $\overline{s}(\phi^+)$ is stable.

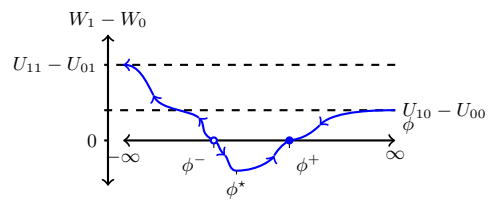

Figure 2: $W_1(\phi) - W_0(\phi)$ (blue line) is a strictly quasi-convex (concave) function of ϕ. $\phi^\star$ denotes the unique local extremum. The direction of the arrows is a consequence of Theorem 8 and Corollary 2.1.

Theorem 12. *If the system asymptotically approaches internal equilibrium, the nullity of $\|D[t]\|_p, p \geq 1$ is preserved in the limit of an infinite time horizon.*

$$\lim_{t' \to \infty} (W_1 - W_0) = 0 \implies \left(\left(\sum_{g=1}^{n-1} \left| \delta(g, g+1)[t] \right| \right)^{1/p} = 0 \iff \lim_{t' \to \infty} \left(\sum_{g=1}^{n-1} \left| \delta(g, g+1)[t'] \right| \right)^{1/p} = 0 \right) \tag{20}$$

Theorem 12 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 8.

3.2 Stability

Using **linear stability analysis**, 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 10 and Eq. (8), 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.

$$W_{\text{eq}} \stackrel{\text{def}}{=} W_0 \Big|_{\text{eq}} = W_1 \Big|_{\text{eq}} = \overline{W}_g \Big|_{\text{eq}} \quad \forall g \in \mathcal{G} \tag{21}$$

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:

$$J \stackrel{\text{def}}{=} \left[\frac{\partial \mathbf{r}}{\partial \delta(1,2)} \quad \frac{\partial \mathbf{r}}{\partial \delta(2,3)} \cdots \quad \frac{\partial \mathbf{r}}{\partial \delta(n-1,n)} \quad \frac{\partial \mathbf{r}}{\partial \overline{s}} \right] \tag{22}$$

where $\mathbf{r}$ is interpreted as a column vector.

Theorem 13. *The Jacobian J simplifies to a scalar multiplied by a matrix with a single non-zero column $\mathbf{v}$ in the last position.*

$$J \Big|_{\text{eq}} = \frac{1}{W_{\text{eq}}} \left(\frac{d\phi}{d\overline{s}} \right) \left(\frac{d}{d\phi} (W_1 - W_0) \right) \left[\mathbf{0}^{(n \times n-1)} \middle| \mathbf{v} \right], \quad \mathbf{v} \stackrel{\text{def}}{=} \begin{bmatrix} \delta(1,2)(1 - s_1 - s_2) \\ \delta(2,3)(1 - s_2 - s_3) \\ \cdots \\ \delta(n-1,n)(1 - s_{n-1} - s_n) \\ \sum_{g \in \mathcal{G}} \mu_g s_g (1 - s_g) \end{bmatrix} \tag{23}$$

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

Corollary 13.1. *Any vector in the vector space generated by D is an eigenvector of J with eigenvalue 0, and $\mathbf{v}$ is an eigenvector of J with eigenvalue λ at equilibrium:*

$$\lambda \stackrel{\text{def}}{=} \left(\sum_{g \in \mathcal{G}} \mu_g s_g (1 - s_g) \right) \frac{1}{W_{\text{eq}}} \left(\frac{d\phi}{d\overline{s}} \right) \left(\frac{d}{d\phi} (W_1 - W_0) \right) \Bigg|_{\text{eq}} \tag{24}$$

Altering any combination of coordinates in D while leaving $\overline{s}$ fixed corresponds to motion along the equilibrium hyperplane, which occurs in neutral equilibrium as a result of the null eigenvalues for D. The equilibrium is stable to perturbations normal to the hyperplane if and only if λ is negative (and, in discrete-time, greater than -2, to prevent repeated over-corrections) [24]:

Corollary 13.2. *As a consequence of Corollary 2.1, the eigenvalue λ is negative, (and the associated equilibrium hyperplane stable) iff $\frac{d}{d\phi}(W_1 - W_0)\Big|_{\text{eq}} > 0$. This prescribes precisely the value ϕ^+ for the stable equilibrium hyperplane.*

4 Interventions

We now restrict our attention to fairness interventions, focusing first on perturbative interventions near the ϕ^+-hyperplane.

Theorem 14. *By direct consequence of Theorem 4, our assumptions automatically satisfy* ***Equality of Opportunity*** *[16], a fairness intervention that requires of the classifier's policy π*

$$\forall g, h \in \mathcal{G}. \quad \Pr(\hat{Y} = z \mid Y = y, G = g) = \Pr(\hat{Y} = z \mid Y = y, G = h) \tag{25}$$

Theorem 14 indicates that equality of opportunity is completely ineffectual with respect to changing underlying qualification disparities in our model.

Theorem 15. *To first-order approximation, perturbation of ϕ induces motion parallel to* $\mathbf{v}$.

As a consequence of Theorem 15, the qualification distance between any two groups cannot be improved by a **global perturbation** to ϕ by relying on linear system response. Rather, the effects of such interventions must rely on the **non-linear response** of the system and are therefore liable to require *large* perturbations to the classifier's threshold. This finding encourages us to consider interventions with group-dependent threshold perturbations. To this end, we hereafter generalize our classifier such that it independently classifies each group according to group-specific threshold. We denote the vector of these thresholds as $\Phi \stackrel{\text{def}}{=} (\phi_1, \phi_2, ..., \phi_n)$ and assume that, prior to some perturbative intervention, $\phi_g = \phi$ for each $g \in \mathcal{G}$.

Theorem 16. ***Demographic Parity*** *[31], a fairness intervention that requires of the classifier*

$$\forall g, h \in \mathcal{G}. \quad \Pr(\hat{Y} = 1 \mid G = g) = \Pr(\hat{Y} = 1 \mid G = h) \tag{26}$$

requires sign-heterogeneous, group-dependent perturbations to the policy threshold when π is non-trivial.

Unfortunately, satisfying demographic parity is a **non-convex** optimization problem [29] and in our case requires solution of a differential dependent on q_y (see Appendix). We therefore rely on simulation to evaluate this intervention in our model.

Feedback control Arbitrary state transitions in the equilibrium hyperplane may be effected by *group-dependent* perturbations to Φ relying on linear system response. Specifically, to diminish a specific qualification distance $\delta(g, g+1)$ for given g, Φ may be perturbed by a vector quantity $\Delta_g \Phi$, obtained by solving for the linearized system response at equilibrium.

Theorem 17. *At internal equilibrium, infinitesimal perturbation of Φ by*

$$\Delta_g \Phi \stackrel{\text{def}}{=} \eta_g \Big(\frac{\alpha_g}{s_1(1-s_1)}, ..., \frac{\alpha_g}{s_g(1-s_g)}, \frac{\beta_g}{s_{g+1}(1-s_{g+1})}, ..., \frac{\beta_g}{s_n(1-s_n)} \Big) \tag{27a}$$

$$\begin{aligned} \alpha_g &\stackrel{\text{def}}{=} (\mu_{g+1} + \mu_{g+2} + ... + \mu_n), & \beta_g &\stackrel{\text{def}}{=} -(\mu_1 + \mu_2 + ... + \mu_g) \\ \eta_g &= (-\epsilon)\delta(g, g+1), & \epsilon &> 0 \end{aligned} \tag{27b}$$

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

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 $\Delta\Phi \stackrel{\text{def}}{=} \sum_{g \in \mathcal{G}} \Delta_g \Phi = (\Delta\phi_1, \Delta\phi_2, ..., \Delta\phi_n)$. Component-wise, $\Delta\Phi$ is given by

$$\Delta\phi_g = \frac{-\epsilon}{s_g(1-s_g)} \Bigg(\sum_{h=g}^{n-1} \delta(h, h+1) \sum_{i=h+1}^{n} \mu_i - \sum_{h=1}^{g-1} \delta(h, h+1) \sum_{i=1}^{h} \mu_i \Bigg) \tag{28}$$

We refer to this mechanism of intervention as an instance of **feedback control**, noting that it is the optimal for the **linear approximation** of the system dynamics at equilibrium. As a notable computation advantage over demographic parity, the value of the control depends only on the known constants μ_g and the current state of the system $\mathbf{s}$. Rather than regularized loss, the strength of the feedback control can be varied explicitly by setting the **strength parameter** ϵ. In practice, we observe more rapid convergence to equal qualification rates with higher values of ϵ. Finally, we remark that it may be possible to compose group-dependent perturbations to Φ with global perturbations of ϕ so as to avoid directly *penalizing* any group, relying on subsidy alone to effect change in the system.

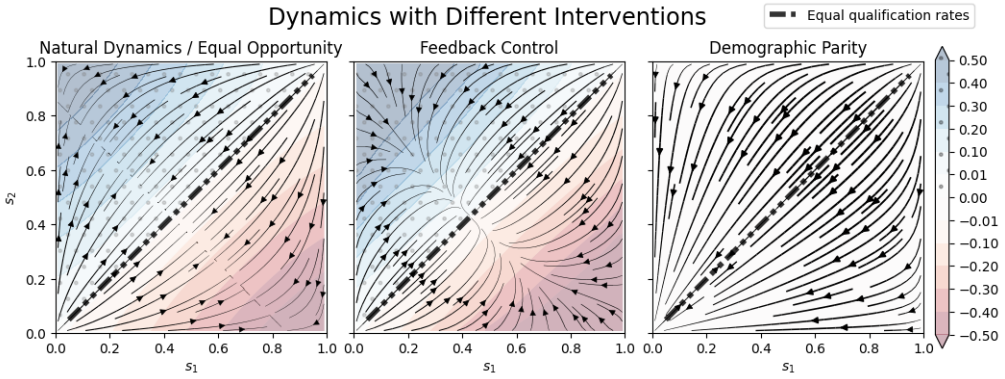

Figure 3: Simulated dynamics for two groups of equal size, subject to different interventions applied globally. $U = [[0.1, 5.5], [0.5, 1.0]]$. $V = [[0.5, -0.5], [-0.25, 1.0]]$, q_0 and q_1 are Gaussians with unit variance and mean -1 and 1 respectively. The streamlines approximate how the system evolves in time, while the background color diplays divergence from demographic parity: $\Pr(\hat{Y} = 0 \mid G = 1) - \Pr(\hat{Y} = 0 \mid G = 2)$. The dashed line from the lower left to the upper right of each subplot demarcates equal qualification rates

Fig. 3 compares the dynamics of our system free from perturbation (which, as we have seen, is equivalent to implementing equal opportunity) to the dynamics of the same system subject to our proposed feedback control and demographic parity. We find that the feedback control mechanism improves convergence towards equal qualification rates even for states far from internal equilibrium. The salient effect of demographic parity for our parameters is to force the system towards a trivial equilibrium at which both groups are unqualified. Subjecting the classifier to demographic parity indeed eliminating disparities, but in a way that we deem socially irresponsible.

5 Discussion and limitations

To the extent that the dynamics of social disparity is not exclusive to algorithmic classifiers [18], efforts to address this problem allow us to reinterpret existing and historical allocation problems, independent of machine automation, with new insight. The novelty of our contribution is the demonstration of persistent disparities in a domain-agnostic setting without assuming structural inequalities. On the contrary, we have made assumptions of structural and cultural homogeneity except for a single binary label. We now face the challenge of relaxing these assumptions and considering how any structural inequalities we introduce will respond to the system's future evolution. It may be argued that the most tenuously justified assumption we have made, and the most consequential towards establishing persistent disparities within our system, is that of isolated subpopulations (Assumption 8).

In evaluating interventions, we acknowledge our lack of rigorous proof regarding demographic parity. However, we have provided a compelling example to reconsider its social consequences. We submit that given the many charitable assumptions of our model towards structural equality, any reasonable fairness intervention should succeed in responsibly rectifying disparities here, if anywhere. We have proposed an explicit feedback control mechanism that succeeds for internal equilibrium states. We leave untreated, however, the possible violations of our fair, underlying assumptions that this intervention might have by breaking group symmetries.

Finally, we invite researchers to consider both our model and explicit control theory applied to society through machine learning, carefully. While potentially preferable to the unconscious selective pressures of current machine classifiers, active regulation of society also raises ethical concerns.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

[2] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *arXiv preprint arXiv:1607.06520*, 2016.

[3] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[4] 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.

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

[6] 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.

[7] Sam Corbett-Davies, Emma Pierson, Avi Feller, and Sharad Goel. A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear. In *Washington Post*, 2016. `https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas`.

[8] Richard Dawkins and Nicola Davis. *The selfish gene*. Macat Library, 2017.

[9] 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.

[10] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, pages 119–133. PMLR, 2018.

[11] 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.

[12] Danielle Ensign, Frielder Sorelle, Neville Scott, Scheidegger Carlos, and Venkatasubramanian Suresh. Decision making with limited feedback. In *Algorithmic Learning Theory*, pages 359–367, 2018.

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

[14] 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.

[15] Swati Gupta and Vijay Kamble. Individual fairness in hindsight. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 805–806. ACM, 2019.

[16] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[17] 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.

[18] 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.

[19] Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Meritocratic fairness for infinite and contextual bandits. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 158–163. ACM, 2018.

[20] 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.

[21] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018.

[22] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3150–3158. PMLR, 2018.

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

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

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

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

[27] Karl Tuyls, Pieter Jan'T Hoen, and Bram Vanschoenwinkel. An evolutionary dynamical analysis of multi-agent learning in iterated games. *Autonomous Agents and Multi-Agent Systems*, 12(1):115–153, 2006.

[28] 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.

[29] 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.

[30] 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, Red Hook, NY, USA, 2017. Curran Associates Inc.

[31] 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.

[32] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827. PMLR, 2013.

[33] 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.

[34] 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.

[35] Daniel Zohary. Unconscious selection and the evolution of domesticated plants. *Economic botany*, 58(1):5–10, 2004.

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] In the supplementary material
   (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 supplementary material
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]