Fairness and Signaling in Bargaining Games

Mihaela Popa-Wyatt, Roland Muhlenbernd, Jeremy L. Wyatt, Cailin O'Connor

Abstract

Cultural evolutionary models of bargaining can elucidate issues related to fairness and justice, and especially how fair and unfair conventions and norms might arise in human societies. One line of this research shows how the presence of social categories in such models creates inequitable equilibria that are not possible in models without social categories. This is taken to help explain why in human groups with social categories, inequity is the rule rather than the exception. But in previous models, it is typically assumed that these categories are rigid—in the sense that they cannot be altered, and easily observable—in the sense that all agents can identify each others' category membership. In reality, social categories are not always so tidy. We introduce evolutionary models where the tags connected with social categories can be flexible, variable, or difficult to observe, i.e., where these tags can carry different amounts of information about group membership. We show how alterations to these tags can undermine the stability of unfair conventions. We argue that these results can inform projects intended to ameliorate inequity, especially projects that seek to alter the properties of tags by promoting experimentation, imitation, and play with identity markers.

1 Introduction

Philosophers and economists use models of bargaining games to understand issues related to fairness and justice, and especially how fair and unfair conventions and norms might arise in human societies.[1] One line of this research shows how the addition of social categories—like racial and gender groups—to bargaining models deeply impacts outcomes (Axtell et al., 2007; O'Connor, 2019; Bruner, 2019). The presence of such categories allows for inequitable equilibria that are not possible in models without them. At these equilibria one social group gets more and another gets less, but there is no particular justification for this pattern beyond the fact that they are part of different social groups. In such models, category markers carry information between interactive partners that facilitate unfair rules like "men get more and women get less". This fact is taken to help explain why in human groups categorical inequity is the rule rather than the exception.

---

[1] See, for example, Skyrms (1994, 2014); Alexander & Skyrms (1999); Alexander (2000, 2007); Bruner (2019); O'Connor (2019); Binmore (1994b,a, 2014); Rubin & O'Connor (2018); H. P. Young (1993b,a); Axtell et al. (2007).

But in this literature it is typically assumed that these categories are rigid—in the sense that they cannot be altered, and easily observable—in the sense that all agents can identify each others' category membership.[2] In reality, social categories are not always so tidy. Sometimes people can voluntarily adopt tags that signal membership in one social category or another. Sometimes social category markers are hard to read, so that interactive partners struggle to reliably identify categorical membership. Sometimes identity markers come in degrees, rather than tidy buckets. In all these cases, the information content of tags or identity markers may be imperfect in ways that disrupt inequitable patterns. If it is not possible to easily tell who is part of which gender, for example, it is not possible to develop a gender rule like "men get more".

We introduce a series of cultural evolutionary models where the markers or tags connected with social categories can be adaptable or undependable in various ways. We explore how and when alterations to these tags undermine the stability of unfair conventions. In our first model, rather than assuming that tags are inalterable, we allow agents to adopt new tags over cultural evolutionary time. In the second model, we assume that tags are more or less observable such that when agents attempt to identify the group membership of an interactive partner they do not always succeed. The final model, following Bruner (2015), assumes that tags come in degrees—as with age or skin color—and that agents may partially alter their tags. In each case, we see that when the information content of tags is disrupted, i.e., when tags do not reliably identify group membership, inequity is disrupted as well. Throughout this exploration, we consider the role of power in these cultural processes, as it is a key factor shaping how inequity evolves in this kind of model and in the world (Bruner & O'Connor, 2017; LaCroix & O'Connor, 2020; Bright et al., 2023; O'Connor et al., 2019).

We argue that these results can inform projects intended to ameliorate inequity, especially projects that seek to alter the properties of category markers. It has been widely noted that attempts to eliminate or abolish categories like race and gender run into problems related to redressing the effects of historical injustices (Fraser, 1995; Bonilla-Silva, 2003; Brown et al., 2003). But at the same time, many have been tempted by the promise of a world where gender, race, and similar categories cannot underpin inequitable systems because they do not exist in their current forms. Our models point to a middle ground—even modest changes to the expression of traditional social identities may weaken their information content, while still preserving them. This observation supports ameliorative proposals like that from Appiah (1996) and Haslanger (2012) who advocate weakening gender and ethnic identities without fully eliminating them. We point to recent changes

---

[2] There are some exceptions, including O'Connor et al. (2019) who consider intersectional markers, Saunders (2022a,b) who considers the evolution of tags themselves, and models like that from Bruner (2015) where tags come in degrees. In service of quite different research questions, authors like Smaldino et al. (2018); Smaldino & Turner (2022) have considered the role of flexible tags in in-group signaling to facilitate cooperation. This signaling allows individuals to selectively reveal their affiliations and avoid detection from potentially hostile out-groups. Such signaling often arises in environments with power imbalances or systemic inequities where openly revealing one's identity is risky. This line of research is also related to earlier work in economics on "secret handshakes" in cooperation, though there the focus is not solely on identity markers as signs of cooperative intent (Robson, 1990).

in gender systems as a successful example of how to weaken the information content in social category markers.

The paper will proceed as follows. Section 2 discusses relevant previous literature and introduces the sort of model we use here. In section 3 we introduce the notion of identity markers acting in various ways that might transfer only partial information about group membership. The next three sections—4, 5, and 6—present the three models described above. In section 7, we discuss what these models can tell us about ameliorative projects. Section 8 briefly concludes.

2 Game Theory, Bargaining, and Unfairness

As noted, a tradition in philosophy and economics uses game theoretic models to reason about justice and fairness.[3] A number of authors have focused on cultural evolution as the place where norms and conventions of fairness typically arise, and, in particular, have employed evolutionary bargaining models to explain the prevalence of fairness norms in human societies (Sugden, 1986; H. P. Young, 1993a; Skyrms, 1994; Alexander & Skyrms, 1999; Skyrms, 2014; Alexander, 2000, 2007). These authors use the Nash demand game, which will be introduced shortly, as their model of bargaining. Across these models, bargaining groups tend to naturally evolve to make equal bargaining demands of each other, and this is taken to explain the cultural evolution of fairness.

Subsequent authors have used similar models to address the other side of the coin—inequity in human societies. It has been widely documented that in spite of the presence of fairness norms most societies have stable, widespread patterns of inequity (Pateman, 1988; Mills, 1997; Tilly, 1998). And, in particular, many of these inequitable patterns build on categorical differences, such as racial or gender differences (Ridgeway, 2011).

Axtell et al. (2007) develop an early model of this sort. They model a population playing a Nash demand game and learning how to bargain based on their experiences with interactive partners. In one version of the model, all agents interact symmetrically and actors learn to make fair (equal) demands of each other. In another version, they add two irrelevant tags—call these “red” and “blue”—to their agents. These tags are observable but otherwise meaningless markers that agents can use to condition their strategic behavior by treating reds one way and blues another.

Under this small alteration, the outcomes of the model are dramatically altered, such that during intergroup bargaining agents of one type often evolve to systematically get more. This is possible because otherwise meaningless tags transfer information between agents. Observing that an interactive partner is a “blue”, say, can allow agents to coordinate on an inequitable rule like “blues get more and reds less”. Thus, this model and similar variants may help explain the endogenous emergence of inequitable or “discriminatory” conventions across social categories (H. P. Young, 1993b; Hoffmann, 2006; Bowles & Naidu, 2006; Henrich & Boyd, 2008; Stewart, 2010; Poza et al., 2011; Rubin & O’Connor,

---

[3] For example, Binmore (2005) uses games to explain how fairness norms arise in an attempt to resolve differences between the egalitarian (Rawls, 1999) and utilitarian (Harsanyi, 1977) theories of social justice.

2018; Bruner, 2019; Bruner & O'Connor, 2017; O'Connor, 2017; O'Connor, 2019; Cochran & O'Connor, 2019; O'Connor et al., 2019; LaCroix & O'Connor, 2020; Bright et al., 2023; Saunders, 2022a,b; Heydari Fard, 2022; Amadae & Watts, 2023).

It is this type of model we build off for the rest of the paper, so in the rest of this section we introduce it in more detail. The Nash demand game assumes two agents divide a resource of some set value. In its original formulation Nash's bargaining problem allowed players to demand any amount of this resource but we, following previous authors, will look at a simplified "mini-game" to make evolutionary analysis tractable.[4] When two agents (1 and 2) interact, each makes one of three demands or bids: low (L), medium (M), or high (H). For each pair of demands, B_1 and B_2, both agents receive each a payoff u_1 and u_2, respectively. If $B_1 + B_2 \leq T$, where T is the total resource, then the payoff for each agent equals their demand, i.e. $u_1 = B_1$ and $u_2 = B_2$. If $B_1 + B_2 > T$, then each agent i receives a payoff equal to their disagreement point d_i, i.e. $u_i = d_i$. We constrain the disagreement point so that it is less than the low demand L. In other words, the agents split a resource, receive what they request when they make compatible demands, but if they make overly aggressive, incompatible demands they get a lower, disagreement payoff. This type of strategic scenario is widespread in human groups and has been taken to represent a wide range of interactions where humans bargain or otherwise divide resources.

In what follows, we will generally assume that $T = 10$ and the three possible demands are $L = 4$, $M = 5$, and $H = 6$. The disagreement points will vary. This game is shown in Table 1.

	L	M	H
L	$4 , 4$	$4 , 5$	$\boxed{4 , 6}$
M	$5 , 4$	$\boxed{5 , 5}$	d_1 , d_2
H	$\boxed{6 , 4}$	d_1 , d_2	d_1 , d_2

Table 1: A three-strategy Nash demand game, where $T = 10$, $H = 6$, $M = 5$, $L = 4$, and $d_1, d_2 < L = 4$. The payoffs for the row player come first and the column player second.

This model has three pure strategy *Nash equilibria* (boxed in Table 1). These are pairings of strategies where neither actor can switch and yield a better payoff. Because there is no incentive to change strategies, Nash equilibria are often thought of as good predictions for how real agents will act in analogous strategic scenarios. Notice these are the three outcomes where actors perfectly divide the resource—either equally, or else favoring one of the two players. Thus, a general prediction of this game is that humans will fully split resources, but that there are multiple ways to do so and that these options are more or less favorable to each player.

In the evolutionary models of justice described above, populations culturally evolving to play this game tend to end up at the equilibrium where the entire group demands M.

---

[4] See J. F. Nash (1950) for his original work on bargaining.

This is the only symmetric equilibria, and thus the only one that an entire, identical group can settle on and always coordinate.[5]

In models with two groups, on the other hand—i.e., where actors have two visible tags—all three Nash equilibria are (typically) stable evolutionary endpoints between the groups. At these outcomes, one group demands H and the other L; they both demand M; or the first L and the other H.[6] The unfair outcomes arise commonly under a variety of assumptions about how individuals learn or culturally evolve, and are thus taken to help explain unfairness in the wild.

2.1 Power and Evolutionary Bargaining

As mentioned, power will play an important role in our analysis below. We address power here because it is deeply important in shaping bargaining conventions between groups (Ridgeway, 2011) and will be relevant to exploring how and where the information in tags can ground inequity. The concept of power is multifaceted and there are multiple ways to operationalize it in bargaining games. Here, we draw on early work by J. Nash (1953) who pointed out that differences in disagreement points between two actors can capture power differences. He focused on an interpretation where these were shaped by threats—each player could make a threat of how they would harm their opponent should bargaining break down. Disagreement points have also been used to track economic, social, or political differences that shape the fall-back positions bargainers face if they fail to reach agreement, including differences of this sort resulting from group identity (Manser & Brown, 1980; McElroy & Horney, 1981).[7]

Bruner & O'Connor (2017) consider two groups evolving to bargain where one has a higher disagreement point than the other. (In the Nash demand game in table 1 this would translate to setting $d_1 > d_2$.) They find that this systematically advantages the empowered group in that the population tends to end up at equilibria where they get more. The greater the power, the stronger the effect.[8] In figure 1 we replicate their results. In the proceeding, we will refer to the two groups as "blue" and "red". We hold the disagreement point for red, $d_R = 0$, and vary the disagreement point for blue, $d_B \in [0, 4]$. Traces show the probability that each possible equilibrium emerges for different power levels: the blues

---

[5] There is another stable outcome involving a mix of H and L demands, but it arises less commonly and is less efficient.

[6] These equilibria correspond to evolutionarily stable states (Maynard Smith & Price, 1973; Selten, 1980). In a nutshell, a population state x is evolutionarily stable if it has an invasion barrier under evolutionary dynamics such that it can repel any invaders (i.e. for small changes away from the state x, the dynamics move back to the state x). These correspond to stable endpoints of many evolutionary dynamics in population-based models.

[7] In this way, the disagreement point can translate to a notion of *power-over* in the sense of imposing one's will on other agents (Weber, 1978; Lukes, 2004), and also the notion of *power-to* act as one wills in the world (Arendt, 1970).

[8] They show this for homogenous groups using the replicator dynamics to represent imitation learning with in the group. LaCroix & O'Connor (2020) replicate their results in an agent-based model with heterogenous power.

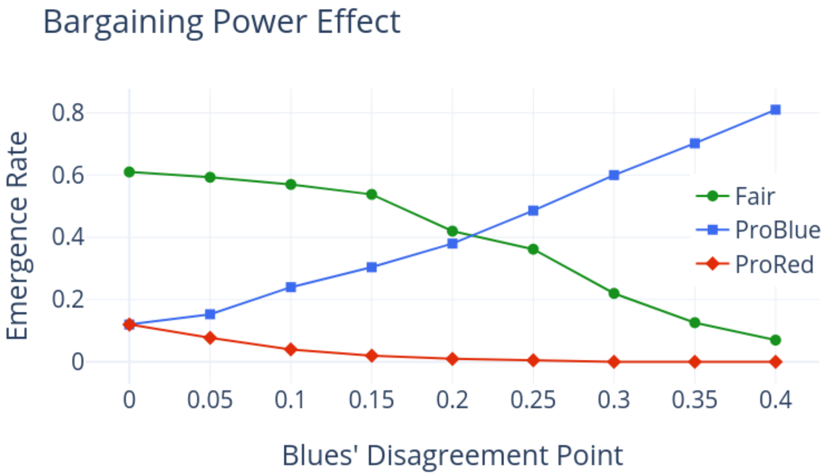


Figure 1: A group with a higher disagreement point will tend to reach favorable bargaining outcomes as a result of cultural evolution. Traces show the emergence rate for three equilibria for different disagreement points of blues ($d_R = 0$; $0 \leq d_B \leq 4$).

demand high and the reds low (proBlue), the blues demand low and the reds high (proRed), or they both make medium demands (fair). As is evident, the greater the power for the blues, the greater the likelihood they end up demanding high, and the less likely they demand low or medium.[9]

3 Groups and Variable Tags

Most previous models in this literature have assumed that tags—the visible markers actors use to condition behavior—are perfectly correlated with group membership. This boils down to an assumption that actors can simply observe group membership and use it to choose strategies. We now move to a more general model where 1) actors belong to underlying social groups, 2) tags are somehow correlated with these group identities, and 3) empowerment is potentially also correlated with group identity. Figure 2 shows this structure.

The reason this general model is important has to do with the conditions under which tags can underpin inequitable equilibria in these models. O'Connor (2019, ch. 9) identifies three conditions that must be in place for inequity to emerge: (i) the presence of categories or tags, (ii) type-conditioning (the ability of agents to condition their strategies on tags of partners), and (iii) learning. She points out that inequity can be disrupted in these models by disrupting any of these three features.

When it comes to disrupting (i), if tags did not exist, inequitable equilibria would disappear. We would revert to the single population models described above. O'Connor

---

[9] These results follow Bruner & O'Connor (2017) in using the two-population replicator dynamics to model cultural change. Each parameter value was simulated 1k times to estimate basins of attraction.

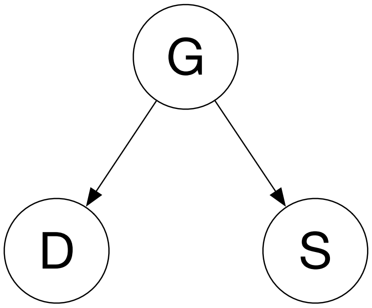


Figure 2: G is fixed, underlying group membership (red or blue). This determines another fixed-trait D (disagreement point) and can influence tag (or signal), S. The arrows represent dependencies, which may be quantified using conditional probability distributions. Note that while tag S is directly observable, group membership G and disagreement point D are not.

points out, though, that some tags are difficult to change—including skin color and some features of biological sex. Others—such as gendered dress, religious dress, and hair color—are easy to change. Still others, like accents, fall in between. Furthermore, different tags may be easier or harder to read as markers of a category. Something like religious dress may be easy to observe and identify compared to some phenotypic features like ambiguous skin color or height. In other words, not all tags are alike in their ability to transfer information about group membership, and not all tags are alike in their flexibility.

In the following, we explore models where the information content in tags is degraded in various ways. As will become clear, we sometimes treat tags as pre-play signals. In a game with pre-play signals, actors first communicate by sending a signal, and then can use these signals to condition strategic behavior in the subsequent game. Pre-play signaling can alter outcomes in evolutionary settings, including in bargaining games.[10] The key difference between signals and tags is that signals are part of a strategy, i.e., they can be adopted and discarded, whereas tags are fixed. But since we explore variations on this theme, below we will sometimes describe tags as signals (i.e., as part of agent strategies).

The previous models we have discussed employ a variety of cultural evolutionary dynamics including ones drawn from evolutionary biology and also agent-based rules for learning and imitation. A general assumption is that via one mechanism or another strategies that are successful tend to proliferate compared to those that are not. Our models use an imitation dynamics called pairwise difference imitation dynamics (PDI), that make the same assumption (Schlag, 1998). Under this rule, success of a strategy translates to probability that it is imitated by a group member. It can be shown that PDI dynamics reproduces features of the replicator dynamics—the most widely used dynamics in evolutionary game theory—in an agent-based context (Schlag, 1998; Izquierdoy et al., 2019).[11]

---

[10] For example, Skyrms (2002) shows that in a single population model pre-play signaling virtually ensures the emergence of fair bargaining conventions.

[11] A Python code implementation of the PDI dynamics and of its application in our three models,

Our simulations thus proceed as follows (unless specified otherwise). Agents of each group are initially assigned a strategy defined as a tuple $\langle S, b_1, b_2 \rangle$, where S is the agent's tag (A or B), b_1 is the demand an agent makes when observing tag A, and b_2 when observing B. In each round, each agent interacts with every other agent, playing based on their strategies, and receiving payoffs. After all play in the round is complete, agents randomly pair with in-group members (blues with blues; reds with reds) and adapt their strategy via probabilistic imitation. The agent with the lower accumulated payoff imitates the agent with the higher payoff with a probability that is proportional to the difference in these payoffs. The simulation continues until the group reaches a stable endpoint or a run-time limit is reached.

In particular, if agents i and j are paired for imitation, and j outperformed i, the probability of imitation is:

$$p_{i,j} = (u_j - u_i)/u_m \tag{1}$$

where u_i is the accumulated payoff of agent i. The denominator, u_m, is the largest possible payoff difference between players.

In many of these models we see endpoints similar to the equilibria described in section 2. (Though details vary based on the model as will become clear below.) In Fair outcomes both groups end up demanding M of each other. In proBlue, when groups meet, blues play H and reds L. ProRed is the exact mirror image. Note that the two discriminatory outcomes (proBlue and proRed) typically correlate with the emergence of a pattern we call distinctive signaling (dS): blues send A, and reds send B.

4 Model 1: Adaptive Tags

In our first model, we assume that tags are more like signals, i.e., that they are sometimes adopted as part of a strategy. In particular, at the beginning of each round of simulation we assume each agent has a small chance (2%) of being paired with another agent from either group to imitate their tag.[12] This means that sometimes red agents can adopt the blue tag, and vice versa. We then simulate bargaining, and the evolution of bargaining and signaling strategies, according to the PDI dynamics.[13] We compare this to a version of the model that is otherwise identical but where tags are held fixed. We vary the power difference between groups, assuming throughout that blues are more powerful than reds ($d_R = 0$ and $0 \leq d_B \leq 4$). For all models presented here, actors start with random bargaining strategies, and with distinctive tags that identify group membership.

Results for both versions of the model are shown in figure 3.[14] In the first condition with no tag imitation (3a), results are very qualitatively similar to those replicated

---

together with an online Appendix, is accessible at:
`https://osf.io/cr3jp/?view_only=a5f1eab56949440e808b6ddadef3802a`

[12] We assume agents always imitate when paired in this way. This feature of the model is thus similar to random mutation of new tags, but limited to only those that exist in the population. We test another version with a 10% chance of tag imitation and get similar qualitative results.

[13] In other words, when agents copy a successful partner, they copy both bargaining strategy and tag.

[14] The population was always 50 reds and 50 blues. For each parameter variation we ran 200 simulations.

above in figure 1. (They vary slightly because we now use PDI dynamics.) As blues become more powerful, they also tend to get more. The results in (3b), though, are quite different.[15] In this version of the model, tags are assigned as before to both groups at the start of simulation, but can be copied as described. Now we see that fair outcomes are much more common for all versions of the model. This is true even without power imbalances, but the effect is especially striking once power imbalances are introduced. In addition, we see that distinctive signaling—each group adopting different tags which identify group membership—is less and less common given power imbalances. The explanation for this effect should be fairly intuitive—if a group is headed toward an inequitable outcome grounded in a tag, the disadvantaged group will learn to switch tags and avoid inequity. The absence of rigid tags thus decreases inequitable outcomes.

We consider two further conditions in which only one group is able to imitate tags, while the other group maintains a constant tag. This amounts to flexibility in signaling for one group, and inflexibility in signaling for the other group. As we will see, this version helps elucidate where flexibility matters. In particular, adaptive tags are helpful for a disempowered group, or any group headed to a disadvantaged outcome. Adaptive tags for a powerful or advantaged group do not disrupt inequity, as such groups are incentivized to maintain distinctive signaling, and thus reach favorable equilibria for themselves.

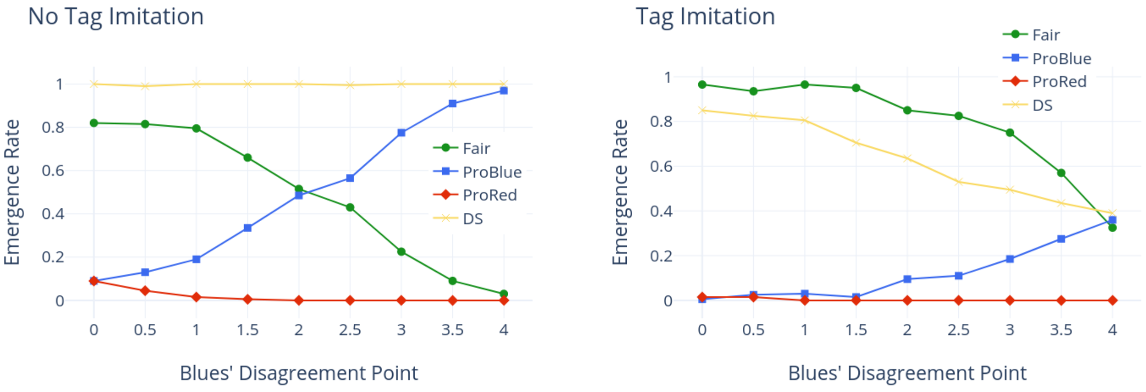
As is clear from figure 4, outcomes are dramatically different in these two conditions. When reds, the disempowered group, are able to change tags, fairness is likely, and distinctive signals are less likely to be maintained. This is because when groups head towards inequitable outcomes, the reds (who are more likely to be disadvantaged) tend to camouflage by adopting the blue group tag. It thus becomes impossible to identify groups, and so impossible for inequity to emerge.

When the blues, the powerful group, have flexible tags, the proBlue outcome tends to emerge. In these cases, the Blues learn to distinguish themselves by adopting distinctive signals. When they do so, the entire system can head towards a proBlue outcome, sup-

---

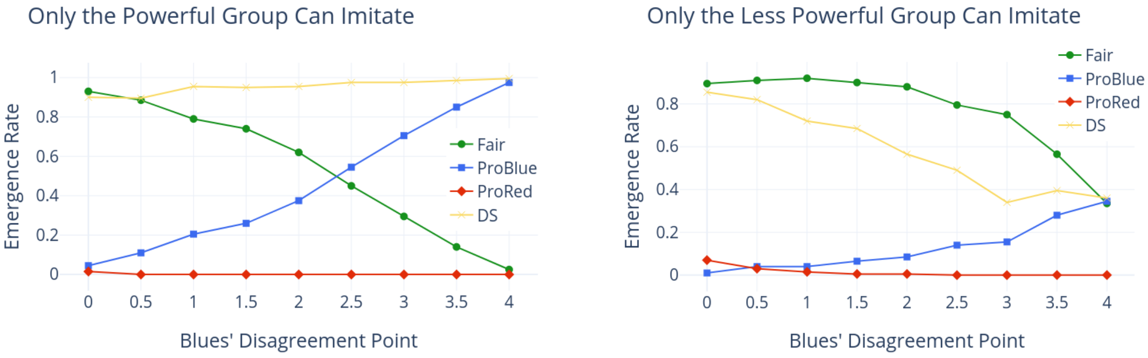
Simulations were halted when agents reached a bargaining equilibrium, an otherwise stable endpoint, or after 1k timesteps. Typically simulations reached a stable endpoint. Some exceptions occurred when $d_B = 4$. Usually simulations would reach an equilibrium state, with some exceptions when a strategy was lost from the population and could no longer be imitated. We report results here for strategies between groups. Within groups, agents typically learned to all play M, and sometimes learned the "fractious" outcome.

[15] The addition of tag imitation complicates outcomes in this model, though we simplify the figures for clarity of communication. In versions with tag imitation, we see outcomes mimicking the equilibria of the base model. We also see outcomes where both groups adopt the same tag, or some mix of both tags, with various behaviors. This often leads to fairness, but there are a few other possibilities. Sometimes we see outcomes we label "aggressive blue". Both groups adopt the same tag, but in response to that tag, blues play H and reds L. This is actually an equilibrium for higher disagreement points, and is reminiscent of the fractious outcome in single population models. This equilibrium is only possible in this model because agents adopt tags from all others, but only copy bargaining strategies from their in-group. We also sometimes see outcomes we label "humble red", where both groups adopt the same tag and in response, blues play M and reds L. This is not an equilibrium, since reds would do better to demand M, but the M strategy is lost from their population. This is somewhat artifactual as the addition of mutation would eliminate this outcome, so we do not focus on it here.



(a) (b)

Figure 3: Tag imitation promotes equity and disrupts the effects of power on bargaining,: 50 blues; 50 reds ($d_R = 0$; $0 \leq d_B \leq 4$), under the PDI dynamics for 200 simulation-runs per data point. dS presents the number of runs with distinct signals across groups.



(a) (b)

Figure 4: If a disempowered or disadvantaged group does not have adaptive signals, inequity is common:($d_R = 0$; $0 \leq d_B \leq 4$), under the PDI dynamics for 100 simulation runs per data point. dS represents the number of runs with distinct signals across groups.

ported by these tags. In other words, when the powerful group adopts a distinctive signal for themselves, they can exploit the inflexible signal of the less powerful group to their advantage.[16]

To summarize, adaptability—the capacity for agents to switch tags strategically—disrupts inequity in these models. This sort of adaptability is especially impactful for disempowered groups.[17]

5 Model 2: Undependable Tags

Our next model considers tags that are undependable, in the sense that they are unreliable or inconsistent in their function or meaning. We can think of these as fixed but stochastic signals. Agents cannot adapt new tags, but instead they signal either A or B with some set probability. Concretely, an agent of group G sends a signal s with probability $p = P(s|G)$, so $p_{A|G}$ is the probability that an agent of G signals A, and $p_{B|G}$ is the probability that an agent of G signals B.

We tested all combinations of values of $p_{.|Blue}$ and $p_{.|Red}$ from the set $[0, 0.25, 0.5, 0.75, 1]$.[18] While the signaling of each agent is fixed during each simulation run, agents adapt their demand strategies according to PDI dynamics, i.e. imitating strategies defined by the tuple $\langle b_1, b_2 \rangle$, where b_1 is the demand upon receiving signal A, and b_2 is the demand upon receiving signal B. We assume throughout that blues have more bargaining power ($d_B = 3$, $d_R = 0$). Results are shown in Fig. 5.

When blues consistently signal A and reds consistently signal B (bottom left corner), or vice versa (top right corner), the proBlue outcome emerges in more than 70% of runs, while fairness emerges in around 20%.[19] This is because tags are perfectly correlated with the types, revealing reliable information about agents' group memberships, and thus creating conditions for blues to exploit their bargaining power. In most of the remaining combinations, fairness emerges with high probability. This shows, again, that distinctive signaling is a precondition for unfairness.

There is, however, a notable asymmetry. Inequity is more possible when reds (the less powerful group) are not perfectly observable. But inequity disappears quickly when the blues (the powerful group) are not perfectly observable. This is somewhat unintuitive, given that in the last section, inequity was more disrupted by flexibility in the signals of the powerless group. Why would this be? In this case, when the blues employ an aggressive

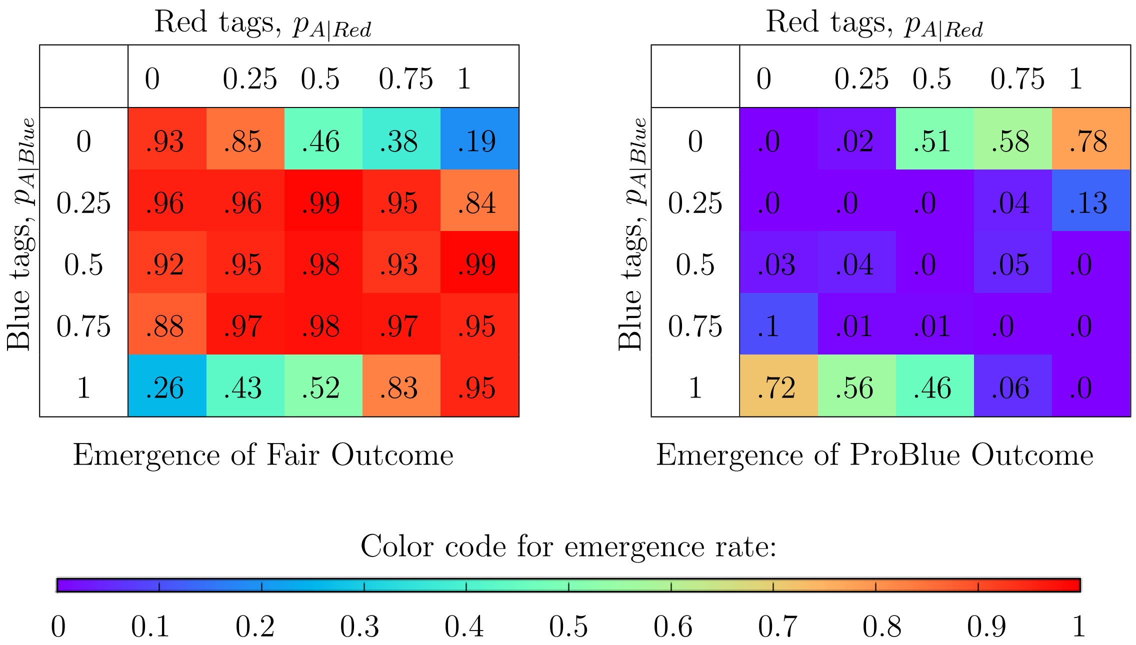
---

[16] Note that in figure 3b, and figure 4b, with enough time, more of the simulations at the proBlue outcome will actually head to the humble red outcome. This will happen as reds copy the blue tag. In other words, our presented results may somewhat overstate the level of inequity that can be stable in these models, which further makes our point.

[17] Also of note: building off the models in this paper, REMOVED FOR REVIEW find that the addition of tag mutation, imitation of group members' tags plus tag mutation, and imitation from out-group members (as we model here), all effectively disrupt inequity, even at low levels.

[18] We once again held the population size at 100, with equal numbers of blues and reds. We ran simulations for 1k timesteps, and ran 100 simulations for each parameter value.

[19] We classified outcomes in the same way described in section 4.



Red tags, $p_{A|Red}$

Blue tags, $p_{A|Blue}$

	0	0.25	0.5	0.75	1
0	.93	.85	.46	.38	.19
0.25	.96	.96	.99	.95	.84
0.5	.92	.95	.98	.93	.99
0.75	.88	.97	.98	.97	.95
1	.26	.43	.52	.83	.95

Emergence of Fair Outcome

Red tags, $p_{A|Red}$

Blue tags, $p_{A|Blue}$

	0	0.25	0.5	0.75	1
0	.0	.02	.51	.58	.78
0.25	.0	.0	.0	.04	.13
0.5	.03	.04	.0	.05	.0
0.75	.1	.01	.01	.0	.0
1	.72	.56	.46	.06	.0

Emergence of ProBlue Outcome

Color code for emergence rate:

Figure 5: This figure shows the emergence rates of fairness (left) and proBlue (right) for different combinations of blues' probability $p_{A|Blue}$ to signal A and reds' probability $p_{A|Red}$ to signal A, where $d_B = 3$; $d_R = 0$. Unfair outcomes are most likely when signaling is distinct.

bargaining strategy, they suffer when they are misperceived. This is because a blue who is making aggressive demands, but is seen as a red, will tend to also meet aggressive demands and reach the disagreement point (both with in- and out-group members). On the other hand, if the blues are making high out-group demands, and reds are misperceived, they all still do well. This points to an interesting possible take-away. In these models, for high power groups it is important that they preserve the dependability of their own tags (more so than preventing lower power agents from using ambiguous tags.)

Still, a low power group may disrupt inequity if their tags are sometimes perceived as out-group. Figure 6 demonstrates this in more detail. We hold fixed the probability of the blue group sending A, $p_{A|Blue} = 1$. We vary the red group's tag dependability from 0 to 1. As is clear, when the powerless group has a less dependable tag, inequity is disrupted.[20]

There is something else to note here, which is that the noise in these models could be produced on either side of an interaction. It may be that some identity signals are inherently ambiguous, or it may be that they are unambiguous, but not dependably observed. This means that if a low power group were able to cultivate indiscriminate observations—by failing to appropriately observe the powerful tag—this could disrupt inequity. Of course,

---

[20] When $p_{A|Red} = 1$ we see fairness drop off slightly. At this parameter setting, both groups are sending the same signal all the time. Some simulations, though, went to the outcome where the blues demand High and the reds Low in response to that signal. This is similar to the fractious outcome in a single group model. It is stable because actors only imitate their in-group.

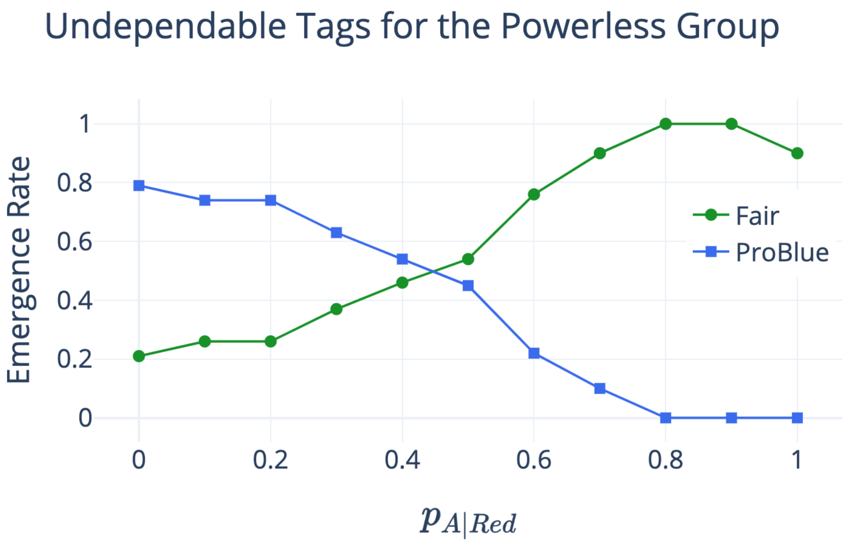


Figure 6: If a disempowered group has an undependable tag, this disrupts inequity. ($d_R = 0$; $d_B = 3$)

in doing so, they would have to temporarily harm themselves, by making over-aggressive demands of their out-group. But the logic is similar to what we see in some cases of social action where oppressed groups refuse to recognize and treat a dominant group as such.

Generally, these results suggest that when inter-group differences are salient and observable, this creates conditions for unfairness to emerge. However, adding even moderate noise—either on the production or perception side—is often enough to reduce the salience of inter-group differences. This, in turn, increases the probability of fairness emerging.

6 Model 3: Continuous (Adaptive) Tags

Our last model is fairly different from those presented thus far. It is intended to give robustness to claims from section 4 using a more realistic and textured set-up. To this point, we have considered binary tags, but in reality tags often vary in more fine-grained ways—skin color, markers of age, some gender markers, and many class markers can come in degrees, for example. In this model, a tag x_i lies in an interval $X = [0, 1]$. The distance between two tags x_i and x_i' is $|x_i - x_i'|$. And, as we will outline in detail shortly, we assume these tags are adaptable.

How does tag conditional behavior work in this model? We assume that every agent has a "tolerance distance" which defines how similar another agent has to be in order to be considered 'in-group' (within the tolerance threshold) or 'out-group' (outside of a tolerance threshold). The strategy of each agent is then a function not of the raw continuous tag, but of whether the tag falls above or below the tolerance threshold. Agents follow a conditional strategy (b_{in}, b_{out}), where b_{in} is the demand they make against in-group, and

b_{out} is the demand they make against out-group members.[21] We assume all agents are randomly given a fixed tolerance threshold drawn from a normal distribution with mean 0.5 and standard deviation 0.1. We considered versions of the model where agents had adaptable tolerance thresholds, but results did not differ substantially. Each agent is also initially randomly assigned a bargaining strategy (b_{in}, b_{out}).

We assume agents may slowly change their tags. Tags have a *degree of alignment* $DA(x)$ which determines how quickly the tag can change under PDI dynamics.[22] Rather than perfectly adopting the tag of a successful imitative partner, an agent will shift some percent of the way, defined by $DA(x)$.[23] Although we experimented with various values of $DA(x)$, it did not end up mattering much to our results, so below we always set $DA(x) = .1$.[24] We assume in-group copying only for both tags and bargaining strategies.

We again assume two underlying groups, blues and reds, which differ in disagreement points ($d_B = 3$; $d_R = 0$). Furthermore, we constrain the simulation such that the two groups begin clustered in different sections of the trait space with blues' tags initialised within the interval [0,0.5] and reds' tags in [0.5,1]. Note that this means that across simulations it is, in principle, possible for there to be small differences in how much the two groups can come to "look" like each other (because we do not allow out-group tag imitation.)[25]

Depending on the details of initial tag distributions, tolerance thresholds, and bargaining strategies, simulations of these models sometimes end up more or less "segregated" with respect to tag values, and with more or less fairness. The main thing we report here is how eventual tag integration shapes the emergence of inequity in these models. The central finding is that when agents adapt tags such that they perceive all or most others as in-group, fairness tends to dominate. It is only in cases where significant in-group/out-group perception remains that we see unfairness.

To quantify this, we measure the ratio of the 'mean tag distance' (MTD) and 'between-types mean tag distance' (BTD). MTD is defined as the mean of distances between two tags of every pair of agents. This tracks how far apart agent tags are on average across

---

[21] We are inspired by Bruner (2015) who explores models where agents of this sort play the stag hunt game. He shows that with evolving tolerance and plastic traits, agents can evolve into a fair and diverse society in which agents cooperate not just with those similar to themselves but also with those of different groups. Using a similar model, Riolo et al. (2001) also show how cooperation can be sustained in donation games.

[22] To be completely clear, when agents engage in imitation according to PDI in these models they imitate bargaining strategy, and also tag according to their DA.

[23] For example, suppose that an agent i intends to imitate a better-scoring agent j. Agent j's tag value is 0.8 and agent i's is 0.1. If $DA = 1$, agent i would adopt agent j's trait value 0.8. But if, for example, DA is 0.5, then agent i would approximate agent j's value half the way, which is $0.1 + (|0.8 - 0.1| \cdot 0.5) = 0.45$.

[24] For greater $DA(x)$ values we see that a global outcome (fair or proBlue) emerges less often. But when it emerges, it patterns in the same way that we report here. For the results of a simulation experiment with $DA(x) = .3$, see Appendix A.3.

[25] If, for example, the highest blue tag is .4 and the lowest red tag is .6, members of the two groups can never be closer than .2.

all agents.[26] BTD is defined as the mean of distances between two tags of every pair of agents, of which one is red, the other is blue. It tracks how far apart the tags of the two groups tend to be on average.[27] Comparing these two measures allow us to say how different tags in the two groups tend to be from each other compared to how different all tags are.

We ran 100 simulations. At the end of each run, we recorded MTD, BTD, and the bargaining outcome.[28] As is clear from figure 7, the relative segregation of the two groups predicts the emergence of unfairness. When there is a lot of distance between the two groups with respect to tags, compared to overall tag distance, we tend to see inequity. When the two groups are relatively integrated, we tend to see fairness.[29]

This model is more complex than the other two we have presented, and there are many ways we might explore it further. What we present here are limited results that lend weight to the findings in section 4. Here, again, we find that the degree to which agents can adapt their tags determines how much inequity tends to arise in our models. And, in particular, in cases where groups more thoroughly mesh identity signals, the less inequity we see.

7 Unfairness, Information, and Identity

Across all three models, there is a unifying theme—when tags are able to reveal information about group membership, and thus about what strategy a partner might play, they can underpin inequitable systems. When the information in tags, about group identity and about the expected behavior of a partner, is decreased, they can no longer support inequity. We give a more in-depth discussion about what it means for a tag to reveal information of this sort, in Appendix A.4. The important thing is that for inequity to work, tags must be able to act as symmetry breakers. They must function to make rules like "blues get more and reds get less" work. When tags are not transferring enough information, these rules cannot get off the ground. A central result here, though, is that the information in

---

[26] The formula for MTD is:

$$MTD = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|$$

[27] The formula for BTD is:

$$BTD = \frac{1}{m \cdot (n-m)}\sum_{i=1}^{m}\sum_{j=m+1}^{n}|x_i - x_j|$$

where n is the total number of agents, agents with index x_1 to x_m are blues, and agents x_{m+1} to x_n are reds.

[28] A bargaining outcome was labelled as fair or proBlue, when more than 2/3 of each group played the appropriate strategy. All other outcomes were labelled as 'other'. Note that this is a less stringent standard than employed elsewhere in the paper, as outcomes in this more complex model were more variable.

[29] As is obvious from this picture, a small number of simulations did not reach one of these outcomes.

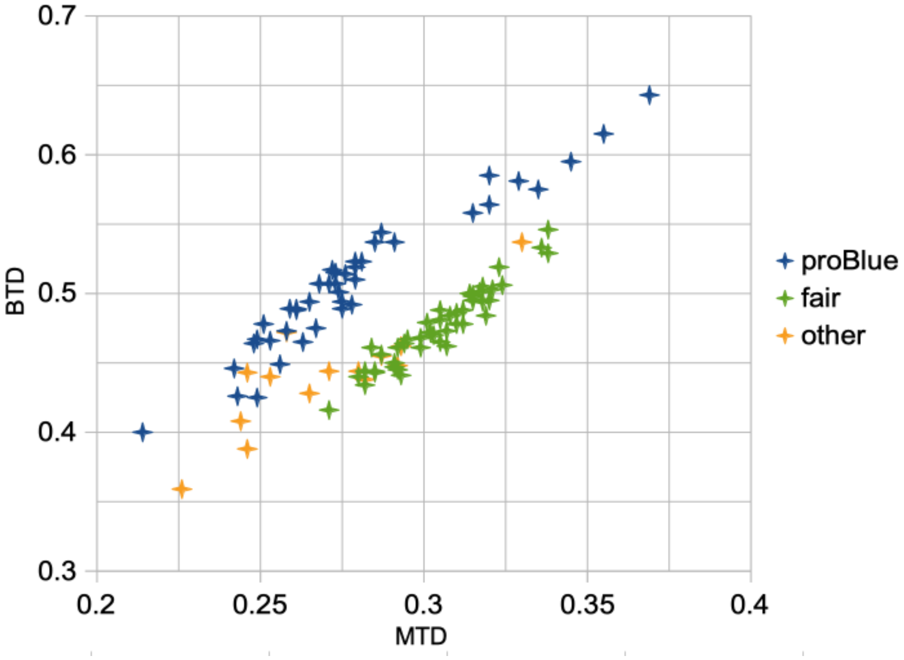


Figure 7: Unfair outcomes are more likely when tags between groups are relatively less similar.

tags need only be decreased to some degree.

One way, then, to disrupt existing inequitable systems is to disrupt the presence of information carrying identity tags. In discussions of social justice or inequity, similar ideas have arisen many times. Some argue for versions of gender or racial "abolition", for example—the idea that these categories should be eliminated, or else seriously reworked, in order to avoid inequity. Okin et al. (1989) gives an early argument for the development of a "genderless" society with "genderless institutions and customs" to prevent gender bias (107). Fraser (1995) suggests a project of undermining gender and racial categories so that, "hierarchical [categories] are replaced by networks of multiple intersecting differences that are demassified and shifting". Haslanger (2012) develops an influential argument for the reworking of gender as we know it in order to avoid the existence of oppressed gender categories. She develops a similar—arguably more radical—argument for racial categories. Others, like Mikkola (2016), push back, arguing that a category like "woman" is not necessarily connected with oppression, and could be maintained while removing inequity. Authors like Escalante (2016), though, are skeptical that such a thing is actually possible. Escalante argues that gender is inextricably linked to power and power differentials, no matter how it is constituted. With respect to projects changing the concept of gender (including those creating more genders), Escalante writes that, "If all of our attempts at positive projects of expansion have fallen short and only snared us in a new set of traps, then there must be another approach... Our only path is that of destruction."[30]

---

[30] We do not include a page number because this is from an online manifesto available at

But there are ethical harms that may arise from the "destruction" of social categories. Haslanger (2012) points out that because biological sex will always be relevant to human cultures, it will not be desirable (and probably not possible) to fully eliminate or destroy categories related to sex. Cull (2019) points to the harms trans people suffer if gender is abolished. If a trans woman strongly identifies as a woman, abolishment of that category is a harm to her.[31] And it has been widely pointed out that for historically marginalized social categories, reworking or eliminating the category might destroy resources necessary for addressing this marginalization. Brown et al. (2003) make this case extensively for race—arguing that a move to "color blindness", or else to eliminating races, ignores the ongoing injustice faced by Black Americans.[32] Furthermore, such a perspective can help dominant, white groups ignore the continuing harms of racism (Gallagher, 2003; Bonilla-Silva, 2003). Even Haslanger (2012), who argues for the elimination of "public or personal identities" based on racial markers, still advocates maintaining a recognition of racial groups to "remedy ongoing injustice" (255).

This tension relates to what Fraser (1995) calls the redistribution-recognition dilemma in thinking about distributive justice. Social justice seems to require that we recognize and attempt to valorize traditionally marginalized groups, while it also demands that we redistribute goods and status. Given what we know about the emergence of inequity, though, the presence of the categories we valorize stands in the way of this sort of redistribution.

How to proceed? We think our models help point towards a path where historically important social categories can be preserved and recognized while simultaneously reducing their ability to support inequity. Our models suggest that even relatively small adjustments to category markers—changes in their dependability, or adaptability—might help disrupt inequitable systems. On this picture, important social categories need not be "destroyed" or even radically restructured. Instead, changes that lower the information content of category markers, while maintaining categories for the most part, may successfully disrupt inequitable systems. This possibility may also dehorn the redistribution-recognition dilemma—if it is possible to recognize social categories, but reduce their importance to the point where equitable distribution is possible, we can have our cake and eat it too. Of course our models are just models, so further empirical support is needed for this possibility. But if we are correct, ameliorative projects that seek to make even modest changes to categories like gender and race may be successful.

Of import here is the fact that social categories and their tags or markers are not, generally, identical. Often a social category might be preserved while its members shift or adapt their tags in ways that decrease the information in these tags. Historically, laws and conventions were adopted for the express purpose of preventing people from doing just this. Sumptuary laws, for example, prevented marginalized racial groups from dressing in

---

https://libcom.org/article/gender-nihilism-anti-manifesto-alyson-escalante

[31] Though Cull's argument makes a substantive empirical assumption—that in a gender-less world, if one is possible, trans identities will be produced in a similar way to our current, very gendered culture. Thanks to REDACTED for this point.

[32] For other critiques of color-blind approaches to racial inequality, see Mills (1997); I. M. Young (1990); Appiah (2005); Anderson (2010).

ways that might confuse their racial identity (Pastore, 2002; Earle, 2003). Such laws help prop up inequitable systems by preventing just the sort of playing with tags and markers we have in mind. But since race is social, not just biological, and since gender is not sex, when such laws are not in place, it will often be possible to play with the markers for these categories.

This proposal is not a radically new one, but rather stands in support of previous thinkers who advocate reducing the power, strength, importance, or salience of social categories without eliminating them altogether. Appiah (1996), for example, suggests that the importance of racial and ethnic identities should be weakened to lessen their power. He writes, “so here are my positive proposals: live with fractured identities; engage in identity play; find solidarity, yes, but recognize contingency” (135). Our models suggest that this might be effective. In addition, they point to specific ways to proceed with “identity play” by highlighting the sorts of changes that reduce the information in category markers.

Our models also support some aspects of gender reworking as proposed by Haslanger (2012). Haslanger argues, “One can encourage the proliferation of sexual and reproductive options without maintaining that we can or should eliminate all social implications of anatomical sex and reproduction” (253). The models here suggest that something along these lines, but probably easier than what she has in mind, might have similarly good outcomes. It may not be necessary to fully revamp gender to prevent it from acting as a locus of oppression. Smaller changes that reduce the information in gender markers may do the work.

Arguably, recent changes in gender expression, expression of sexual orientation, and other aspects related to gendered behavior are disrupting gender categories in the right sorts of ways. O’Connor (2019), using models like the ones developed here, points out that for gender to divide labor, status, and resources, a “bundle” of features must be in place—sex and gender must be strongly correlated, individuals must economically (and usually romantically) pair with those of the “opposite” sex/gender, gender signals or markers must correlate strongly with gender, and individuals must learn and adhere to their proper gendered roles. Changes that disrupt any aspects of this bundle can help disrupt the system. We are focused here on one aspect of this disruption—changes related to gender signals or markers, such as those coming out of the feminist, gay rights, and trans rights movements. When individuals adopt confusing or non-traditional gender signals, when they choose romantic partners in different ways, when they engage in roles and jobs that are not associated with their gender, when they decouple gender presentation from sex—these actions may all help disrupt inequity.

Note that our claim is not that individuals are morally required to make decisions about these issues for the purpose of inequity. Gender and racial expression, for example, are often deeply personal and many report harms from engaging in inauthentic identity expression. In addition, when it comes to expressions that constitute ‘passing’ as part of another social group, there are complex ethical and practical reasons for and against.[33]

---

[33] Silvermint (2018) argues that passing as privileged can lead to inauthentic identity expression, causing both personal and social harm. On a personal level, it can create psychological distress, identity conflict, and feelings of alienation, as individuals must hide their true selves. Socially, passing can uphold oppression

But it seems to be the case for gender that facilitating freedom of expression is, in fact, enough to create the kinds of disruptions we have in mind. In the presence of such freedom, many individuals seem to want to create these disruptions. In other words, simply allowing individuals to play as they like without significant normative push-back may be enough.

As noted, Escalante argues that gender in any form creates harmful power disparities. On this account, disruptions to gender systems like the adoption of "non-binary" as a category, or of trans identities, simply create more loci for inequity. It is ultimately an empirical question, though, whether or not this is true. Throughout much of human history, gender has been produced in a binary system that builds off sex differences. The models in this paper where inequity is easy to produce are those that most neatly track traditional gender systems—everyone is observably part of one category and they cannot change. And, in any of our variants where information transfer is sufficiently lessened, we see inequity reduced. It may be the case that there is hope for this sort of more modest categorical change.

One last note—we argue that modest identity play may help undermine inequitable systems. But highly oppressed groups rarely have the freedom to express their identities as they like. Oppressive systems often include rigid rules and normative punishments for those who attempt to alter their identity markers, in part because by enforcing identity, powerful groups can retain their power. Thus, many inequitable systems cannot be dismantled by alterations to identity markers alone. Instead, changes of many sorts—including to economic, material, and political conditions, as well as to identity expression—may be necessary to promote equity.[34]

8 Conclusion

We present three models where tags are either 1) adaptable or 2) undependable to differing degrees. In all three cases, the information that tags can transfer about an interactive partner is attenuated as a result. And, as we demonstrate, in all three cases, this means that it is harder to culturally evolve inequitable systems. When tags carry a good amount of information about group membership, and about expected bargaining behavior, the chance that other agents exploit this information increases. Conversely, when signals are less informative, this exploitation is prevented and fairness is more likely to emerge.

We argue that this observation may have important consequences for thinking about social justice. It may be possible to preserve historically oppressed social categories, while playing with or modifying the tags or markers associated with these categories, so they carry less information about expected behavior. Our models give some guidance into how this might work—changes intended to make category membership harder to read, or

---

by hiding discrimination and making privilege seem normal or earned. While passing may be necessary for safety or avoiding prejudice, Silvermint also highlights how it can unintentionally support the very systems that marginalize people. In addition, as authors like Smaldino et al. (2018) and Smaldino & Turner (2022) highlight, there can be other strategic reasons why in-group identity signaling is important.

[34] Thanks to REDACTED for this point.

to uncorrelate markers with underlying identities, can make inequity harder to sustain. These sorts of changes encompass the gender play we describe above, but also might involve small changes to class signals, accent, or ethnic identity. As noted, we do not think that individuals should consider it an ethical responsibility to alter or play with their identity markers. But in cases where individuals are inclined to do so even modest play may have beneficial consequences for equity. Alternatively, our models may help explain why oppressive regimes so often police identity signaling—because such signaling is core to propping up inequity.

In addition, changes not to markers, but their observability may be effective for similar reasons. There are many examples of policies intended to reduce observability of markers. Anonymized auditions for orchestras (Persson, 2022) and other anonymized hiring processes——where identifying details such as name, gender, and socioeconomic background are obscured——prevent employers from recognizing identity. Likewise, lotteries for school or other admissions processes can reduce focus on identity markers in decision making (Basteck et al., 2021). Such structural changes are ways to decrease information transfer via identity tags, without putting all the responsibility on individuals, and so are well worth doing.

Of course, the conclusions we draw here are based on highly simplified social models. We take them to be suggestive, to aid normal reasoning about inequity, and to provide concrete suggestions that can be further empirically explored. In particular, we think further study into the degree of information in tags, and how they do or do not support inequity, is worthwhile.

9 Acknowledgements

Thanks to Liam K. Bright for comments on a draft of this paper and Sahar Heydari-Fard for discussion of it. Thanks to audiences at the University of Toronto Simon Lectures for comments.

References

Alexander, J. (2000). Evolutionary explanations of distributive justice. *Philosophy of Science*, *67*(3), 490–516.

Alexander, J. (2007). *The structural evolution of morality.* Cambridge University Press.

Alexander, J., & Skyrms, B. (1999). Bargaining with neighbors: Is justice contagious? *The Journal of philosophy*, *96*(11), 588–598.

Amadae, S., & Watts, C. J. (2023). Red queen and red king effects in cultural agent-based modeling: Hawk dove binary and systemic discrimination. *The Journal of Mathematical Sociology*, *47*(4), 283–310.

Anderson, E. (2010). *The imperative of integration.* Princeton University Press.

Appiah, K. A. (1996). Race, culture, identity: Misunderstood connections. *Tanner Lectures on Human Values*, *17*, 51–136.

Appiah, K. A. (2005). *The ethics of identity.* Princeton University Press.

Arendt, H. (1970). *On violence.* Houghton Mifflin Harcourt.

Axtell, R., Epstein, J., & Young, H. P. (2007). The emergence of classes in a multi-agent bargaining model. In *Generative social science* (pp. 177–195). Princeton University Press.

Basteck, C., Klaus, B., & Kübler, D. (2021). How lotteries in school choice help to level the playing field. *Games and Economic Behavior*, *129*, 198–237.

Binmore, K. (1994a). *Game theory and the social contract: Just playing.* MIT Press. Retrieved from `https://books.google.de/books?id=HZ1hC1MLPeoC`

Binmore, K. (1994b). *Game theory and the social contract, volume 1: Playing fair.* MIT Press. Retrieved from `https://books.google.de/books?id=LkJAxAEACAAJ`

Binmore, K. (2005). *Natural justice.* Oxford university press.

Binmore, K. (2014). Bargaining and fairness. *Proceedings of the National Academy of Sciences*, *111*(Supplement 3), 10785–10788.

Bonilla-Silva, E. (2003). *Racism without racists: Color-blind racism and the persistence of racial inequality in america.* Rowman & Littlefield.

Bowles, S., & Naidu, S. (2006). Persistent institutions. *Working paper*(08-04).

Bright, L. K., Gabriel, N., O'Connor, C., & Taiwo, O. (2023). On the stability of racial capitalism. *Ergo*.

Brown, M. K., Carnoy, M., Currie, E., Duster, T., Oppenheimer, D. B., Shultz, M. M., & Wellman, D. (2003). *Whitewashing race: The myth of a color-blind society.* Univ of California Press.

Bruner, J. (2015). Diversity, tolerance, and the social contract. *Politics, Philosophy & Economics*, *14*(4), 429–448.

Bruner, J. (2019). Minority (dis) advantage in population games. *Synthese*, *196*(1), 413–427.

Bruner, J., & O'Connor, C. (2017). Power, bargaining, and collaboration. In T. Boyer-Kassem, C. Mayo-Wilson, & M. Weisberg (Eds.), *Scientific collaboration and collective knowledge.* Oxford University Press.

Cochran, C., & O'Connor, C. (2019). Inequality and inequity in the emergence of conventions. *Politics, Philosophy & Economics*, *18*(3), 264–281.

Cull, M. J. (2019). Against abolition. *Feminist Philosophy Quarterly*, *5*(3).

Earle, R. (2003). Luxury, clothing and race in colonial spanish america. In M. Berg & E. Eger (Eds.), *Luxury in the eighteenth century: Debates, desires and delectable goods* (pp. 219–227). London: Palgrave Macmillan UK. Retrieved from `https://doi.org/10.1057/9780230508279_16` doi: 10.1057/9780230508279_16

Escalante, A. (2016). Gender nihilism: An anti-manifesto. *Libcom. Org.*

Fraser, N. (1995). From redistribution to recognition? dilemmas of justice in a'post-socialist'age. *New left review*(212), 68.

Gallagher, C. A. (2003). Color-blind privilege: The social and political functions of erasing the color line in post race america. *Race, Gender & Class*, 22–37.

Harsanyi, J. C. (1977). Morality and the theory of rational behavior. *Social Research*, *44*(4), 623–656.

Haslanger, S. (2012). *Resisting reality: Social construction and social critique.* Oxford University Press.

Henrich, J., & Boyd, R. (2008). Division of labor, economic specialization, and the evolution of social stratification. *Current Anthropology*, *49*(4), 715–724.

Heydari Fard, S. (2022). Strategic injustice, dynamic network formation, and social movements. *Synthese*, *200*(5), 392.

Hoffmann, R. (2006). The cognitive origins of social stratification. *Computational Economics*, *28*(3), 233–249.

Izquierdoy, L. R., Izquierdoz, S. S., & Sandholm, W. H. (2019). An introduction to ABED: Agent-based simulation of evolutionary game dynamics. *Games and Economic Behavior*, *118*, 434–462.

LaCroix, T., & O'Connor, C. (2020). Power by association. *Ergo*.

Lukes, S. (2004). *Power: A radical view.* Macmillan International Higher Education.

Manser, M., & Brown, M. (1980). Marriage and household decision-making: A bargaining analysis. *International economic review*, 31–44.

Maynard Smith, J., & Price, G. R. (1973). The logic of animal conflict. *Nature*, *246*(5427), 15–18.

McElroy, M. B., & Horney, M. J. (1981). Nash-bargained household decisions: Toward a generalization of the theory of demand. *International economic review*, 333–349.

Mikkola, M. (2016). *The wrong of injustice: Dehumanization and its role in feminist philosophy.* Oxford University Press.

Mills, C. W. (1997). *The racial contract.* Cornell University Press.

Nash, J. (1953). Two-person cooperative games. *Econometrica: Journal of the Econometric Society*, 128–140.

Nash, J. F. (1950). The bargaining problem. *Econometrica: Journal of the econometric society*, 155–162.

O'Connor, C. (2019). *The origins of unfairness: Social categories and cultural evolution.* Oxford University Press, USA.

Okin, S. M., et al. (1989). *Justice, gender, and the family* (Vol. 171). Basic books New York.

O'Connor, C. (2017). The cultural red king effect. *The Journal of Mathematical Sociology*, *41*(3), 155–171.

O'Connor, C., Bright, L. K., & Bruner, J. P. (2019). The emergence of intersectional disadvantage. *Social Epistemology*, *33*(1), 23–41.

Pastore, C. (2002). Consumer choices and colonial identity in Saint-Domingue. *French Colonial History*, *2*(1), 77–92.

Pateman, C. (1988). *Sexual contract.* Polity Press.

Persson, A. (2022). *Problems with the veil of ignorance, and how we might solve them.*

Poza, D. J., Villafánez, F. A., Pajares, J., López-Paredes, A., & Hernández, C. (2011). New insights on the emergence of classes model. *Discrete Dynamics in Nature and Society*, *2011*.

Rawls, J. (1999). *A theory of justice: Revised edition.* Harvard university press.

Ridgeway, C. L. (2011). *Framed by gender: How gender inequality persists in the modern world.* Oxford University Press.

Riolo, R. L., Cohen, M. D., & Axelrod, R. (2001). Evolution of cooperation without reciprocity. *Nature*, *414*(6862), 441–443.

Robson, A. J. (1990). Efficiency in evolutionary games: Darwin, nash and the secret handshake. *Journal of theoretical Biology*, *144*(3), 379–396.

Rubin, H., & O'Connor, C. (2018). Discrimination and collaboration in science. *Philosophy of Science*, *85*(3), 380–402.

Saunders, D. (2022a). How to put the cart behind the horse in the cultural evolution of gender. *Philosophy of the Social Sciences*, *52*(1-2), 81–102.

Saunders, D. (2022b). When is similarity-biased social learning adaptively advantageous? *The British Journal for the Philosophy of Science*.

Schlag, K. (1998). Why imitate, and if so, how? a boundedly rational approach to multi-armed bandits. *Journal of Economic Theory*, *78*(1), 130–156.

Selten, R. (1980). A note on evolutionarily stable strategies in asymmetrical animal contests. *Journal of Theoretical Biology*, *84*, 93–101.

Silvermint, D. (2018). Passing as privileged. *Ergo, an Open Access Journal of Philosophy*, *5*.

Skyrms, B. (1994). Sex and justice. *The Journal of philosophy*, *91*(6), 305–320.

Skyrms, B. (2002). Signals, evolution and the explanatory power of transient information. *Philosophy of Science*, *69*(3), 407–428.

Skyrms, B. (2014). *Evolution of the social contract*. Cambridge University Press.

Smaldino, P. E., Flamson, T. J., & McElreath, R. (2018). The evolution of covert signaling. *Scientific reports*, *8*(1), 4905.

Smaldino, P. E., & Turner, M. A. (2022). Covert signaling is an adaptive communication strategy in diverse populations. *Psychological review*, *129*(4), 812.

Stewart, Q. T. (2010). Big bad racists, subtle prejudice and minority victims: An agent-based analysis of the dynamics of racial inequality. In *annual meeting of the population association of america.*

Sugden, R. (1986). The economics of cooperation, rights and welfare. *New York: Blakwell*.

Tilly, C. (1998). *Durable inequality*. University of California Press.

Weber, M. (1978). *The economy and society: an outline of interpretive sociology*. University of California Press.

Young, H. P. (1993a). An evolutionary model of bargaining. *Journal of economic theory*, *59*(1), 145–168.

Young, H. P. (1993b). The evolution of conventions. *Econometrica: Journal of the Econometric Society*, 57–84.

Young, I. M. (1990). *Justice and the politics of difference*. Princeton University Press.