Why Natural Social Contracts are Not Fair

Cailin O'Connor

October 2022

Abstract

Many theorists have employed game theory to model the emergence of stable social norms, or natural "social contracts". One branch of this literature uses bargaining games to show why many societies have norms and rules for fairness. In cultural evolutionary models, fair bargaining emerges endogenously because it is an efficient way to divide resources (Young, 1993a; Skyrms, 1996; Alexander, 2007). In response, a number of authors have argued that these models miss an important element of real human societies – divisions into groups or social categories. Once such groups are added to cultural evolutionary models, fairness is no longer the expected outcome. Instead "discriminatory norms" often emerge where one group systematically gets more when dividing resources (Axtell et al., 2001; O'Connor, 2019). These results may help explain why categorical inequity is the rule across human societies (Mills, 1997; Pateman, 1988). If one wishes to understand the naturalistic emergence of social contracts, one must account for the presence of categorical divisions, and unfairness, as well as for norms of fairness. This paper overviews this body of work, and pulls out lessons for social contract theory.

1 Introduction

There is a rich literature using tools from game theory to reason about, and to model the social contract (Binmore, 1994). One branch of this literature looks at how naturalistic social contracts may have emerged in societies over cultural evolutionary time.[1] Some research along these lines focuses on the establishment of conventions and norms for altruism, cooperation, joint action, and communication using models like the prisoner's dilemma, stag hunt, hawk-dove, and

[1]Mills (1997) draws a helpful divide between social contract theorizing about idealized or justified contracts, and theorizing about the actual history of social contracts. Early work in social contract theory from Hobbes, Locke, Hume, and Rousseau (among others) was mostly in the latter vein. Modern work, notably that in the Rawlsian tradition, is largely about what would make a contract justified. The research reviewed here is about the emergence of actual natural contracts. (So is that from Mills.) Skyrms (1996) draws another helpful distinction among traditional social contract work. Some theorists, notably Hobbes, focus on rational choice as a grounds for the establishment of the social contract. Others like Hume and Rousseau are focused on a process where cultural patterns emerge over time. The work discussed here is in this latter vein.

signaling game (Skyrms, 1996, 2004; Bruner, 2015).[2] In this paper, we discuss work on the emergence of norms governing the division of resources, focusing on bargaining games, and especially the Nash demand game.[3] One especially persuasive branch of argumentation shows how fairness, in the sense of equal divisions of resources, is likely to emerge on cultural evolutionary timescales. Models of this sort are used to argue that natural social contracts often emerge that are, for the most part, fair.

Of course, when we look at real world conventions and norms regarding the division of resources, fairness is not typically the rule. Pateman (1988) in *The Sexual Contract* criticizes traditional social contract theory on the grounds that it fails to recognize how sexism and patriarchy suffuse rules of fairness, equality, and division of resources across societies. Mills (1997) forcefully makes similar points about race in his influential *The Racial Contract*. As he points out, despite the high ideals and optimism of traditional social contract theorists, the real world is rife with inequity. Social contracts often dictate that members of one race will act as a privileged class, while another will be oppressed.

How do we square these observations with the modeling literature showing that fairness emerges naturally via cultural evolution? The answer resides in what is missing from these models. Both Mills and Pateman address inequity that happens across the boundaries of social categories, like race and gender. An emerging literature considers models that include social categories, and shows that this addition opens the door for persistently inequitable natural contracts to emerge. The goal of this paper is to survey this literature, drawing out lessons for social contract theory where applicable.

In section 2 I discuss evolutionary game theoretic modeling showing how and why fair conventions for resource division tend to emerge. The meat of the paper, 3, describes models showing why the presence of social categories shapes cultural evolution such that inequitable norms emerge. This literature addresses a number of relevant topics including power, minority status, intersectional disadvantage, how different sorts of social categories can (or cannot) underpin unfairness, and why social categories emerge in the first place. Section 4 uses this modeling framework to ask: how can we promote fair natural contracts in a diverse world? And section 5 concludes.

Note that this paper is short, and focused on a relatively small area of literature. There are many related topics I might have included: the emergence of cooperation in the stag hunt or prisoner's dilemma, gendered contracts related to division of labor, empirical work in sociology and anthropology on the emergence of unfair norms and conventions, very different approaches to modelling the emergence of inequitable norms etc.[4] The goal here is to focus on one, small

[2] See Binmore in this volume for more on the prisoner's dilemma as a model of the social contract.

[3] See citations herein including Binmore (1994, 2005); Skyrms (1994, 1996, 2004); D'Arms et al. (1998); Alexander and Skyrms (1999); Alexander (2000, 2007). Bruner, in this volume, also focuses on the Nash demand game in simulating what sorts of solutions might be chosen by real-world actors.

[4] For this last, see, for example, Carvalho and Pradelski (2019).

topically and formally unified area of social contract theory.

2 The Emergence of Fair Contracts

Skyrms (1996) shows how fair bargaining conventions can emerge via cultural change and evolution.[5,6] His model makes use of the *Nash demand game*, a simplified representation of scenarios where individuals divide resources.[7] In the model, two agents present demands for some portion of a resource (85%, say, or 42%). If the demands are compatible in that they do not exceed the resource, the two agents receive what they request. If the demands are incompatible in that they over-demand the resource, it is assumed the agents fail to come to an acceptable division. Instead, they each receive a lower payoff, called the *disagreement point*. A game theoretic analysis of this model predicts that agents will settle on splits that perfectly divide the resource. This is because those divisions constitute the *Nash equilibria* of the game—the strategy pairings where neither player is incentivized to change their behavior. At these splits, agents cannot switch to demand more, because doing so will over-demand the resource. If they demand less, they simply get less. Thus these outcomes are expected to be stable.

This model has been used to track many sorts of real-world cases, ranging from international trade agreements to sharecropping contracts to division of household labor. Generally the Nash demand game models scenarios where individuals must divide a resource, where there are many ways they could do it, where these outcomes vary in how favorable they are to the two players, and where too much aggression means that bargaining fails. With respect to social contract theory, they track one of the central problems groups of humans face—who gets how much?

Notice that one of the Nash equilibria of this model is a split that we tend to think of as fair—the 50/50 split. But there are also an infinity of other, unfair equilibria—60/40, and 90/10, and 13/87, for example. The game theoretic prediction seems to be that natural social contracts may take many forms.

There are some crucial aspects missing from this model, though. The analysis just described assumes that agents make rational choices, i.e., that they pick the strategies in Nash demand games that are expected to yield the highest payoffs. The analysis also assumes a scenario where just two agents interact, dividing just one resource. But real social contracts tend to emerge over time, as a result of cultural exchange and learning, rather than just rational choice. And, in addition, they often emerge on a group level in response to many, re-

[5] His broader project involves showing how evolutionary analyses of strategic behavior are often in tension with rational choice analyses. In the case of bargaining, he argues that evolution tends to select fair equilibria out of a plethora of possible equilibria of bargaining games.

[6] For influential work using bargaining models to think about the evolution of fairness norms, see also Binmore (1994, 2005).

[7] Versions of this game go back to John Nash's work on the bargaining problem (Nash, 1950).

peated interactions where different individuals divide resources. For this reason, Skyrms (1996), and following researchers, have focused on *evolutionary game theory* in modeling the emergence of bargaining contracts. In doing so, they reflect social contract theorists like Rousseau and Hume, who recognized that social problems get solved on long, cultural time-scales.

Evolutionary game theoretic models assume that a group of agents play a game repeatedly over time. As they do, they learn or evolve strategies in response to their experiences. A typical assumption, when applying this models in the cultural realm, is that agents will tend to repeat successful strategies, or else to copy strategies that are doing well for others. In this way strategies that do well spread, and those that do poorly die out. Eventually through this process the population will settle into a stable end state which, typically, will correspond to one of the Nash equilibria of the model.

The key finding of this literature is that in evolutionary models, fair outcomes are very common. They tend to emerge endogenously as a result of cultural learning and change. To see how we get there let us look in more detail at the models.

Skyrms (1996), and others, typically have considered simplified versions of the Nash demand game that include only three demands—High, Medium, and Low.[8] This makes evolutionary analysis much more tractable while preserving the basic strategic structure of the game. A payoff table of this sort of simplified game appears in figure 1. Payoffs in this table are shown for any combination of demands, with player 1 listed first. As is evident, this version of the game considers a resource of value 10, with possible demands for 3, 5, or 7. If the players demand more than 10, they get a disagreement point of 0.[9] The three (pure strategy) Nash equilibria are bolded: Low vs High, Medium vs Medium, and High vs Low. In each of these outcomes the resource is perfectly divided, but only one of them—the Medium vs. Medium equilibrium—is usually considered "fair".

		Player 2		
		Low	**Med**	**High**
Player 1	**Low**	3,3	3,5	**3,7**
	Med	5,3	**5,5**	0,0
	High	**7,3**	0,0	0,0

Figure 1: Payoff matrix for a simplified Nash demand game. The pure Nash equilibria are bolded.

Under most standard evolutionary *dynamics* (rules for learning and cultural evolution) a population playing this game is most likely to head towards a sta-

[8] Or else games with another, finite number of strategies, such as 5 or 7 demands as in O'Connor (2019).

[9] We might just as well have picked another version of this game with different values for the demands or the disagreement point. The important thing is the basic structure of the game.

ble outcome where all agents make the Medium demand. In other words, a fair social contract emerges endogenously. The other stable evolutionary outcome is one where most agents make the low demand, and a smaller number make the high one. This is sometimes called a "fractious" outcome, because agents sometimes coordinate their demands, but also sometimes miscoordinate by playing High vs High or Low vs Low.[10] Notice that this outcome is stable because even though actors play different strategies, they all expect the same payoffs. Low demanders coordinate more often, and get less each time. High demanders coordinate less often, but get a higher payoff when they do.[11]

We have two stable outcomes of the evolutionary model, but why is fairness more likely to emerge? The fair outcome is the more efficient one. It is the only population state where any two actors will play a Nash equilibrium any time they interact. This is because Medium vs. Medium is the only symmetric equilibrium. In the fractious outcome, sometimes Lows meet Lows and underdemand the resource, thus wasting it. Sometimes Highs meet Highs, and reach the disagreement point. Because the fair outcome is efficient in this way, actors get higher average payoffs, and for this reason it tends to have more evolutionary pull.

Further work has focused on explaining why fairness might evolve even more often than these initial models would suggest. In general, structural conditions where individuals tend to meet partners who play the same strategies that they do will create greater pressure for fairness to emerge (Skyrms, 1996). Structures with repeated interaction between neighbors tend to select for fairness for this reason (Alexander and Skyrms, 1999; Alexander, 2007; Skyrms, 2004). And communication between neighbors also tends to generate fair outcomes (Skyrms, 2004, 1996).[12]

Thus these models show that fair conventions of behavior do tend to emerge naturally from an uncoordinated "state of nature". They support the idea that natural social contracts tend to favor equality.

3 The Emergence of Unfairness

To this point, I have described how evolutionary game theoretic models support claims about the emergence of fair social contracts. But social contracts are not always fair, as a great deal of research in the social sciences has shown.[13] As mentioned, Pateman (1988) and Mills (1997) have used this point to criticize traditional social contract theory on the grounds that real natural contracts have tended to be deeply unfair.

[10] The exact details of this fractious outcome depend on the payoffs of the game. See Skyrms (1996) for details.

[11] This is inherent in the definition of an evolutionary equilibrium. Otherwise if strategies yield different payoffs, the higher payoff strategy spreads.

[12] In addition, Sugden et al. (1986) shows that the fair outcome is the only evolutionary stable strategy (ESS) of the game, and Young (1993b) that it is the only stochastically stable equilibrium (SSE), lending credence to the idea that it is evolutionarily special.

[13] I do not survey this for space reasons.

How do we square the models presented in section 2 with these criticisms? The answer is that we need to add *social categories* to these models. A social category is a recognizable group within a society. Most important to us here are *primary categories*, which Ridgeway (2011) describes as the small number of social categories most generally used for coordinating behavior. Across societies, these always include gender and age, and often also include race, religion, caste, or class.

We can add social categories to the models described in section 2. Although there are many ways to do this, we will start simple and go from there. Our model will involve a population with two groups (representing social categories) that each have a different arbitrary *tag*. The tags might be "green" and "yellow", for example, or "star-belly" and "plain belly". Agents in this model play the bargaining game shown in figure 1 but in doing so may condition their strategy on the tag of their partner. For example, an agent in the green group might play Medium against other greens, and Low against yellows. We can label this two part strategy, listing the in-group strategy first, as follows: $< \text{Medium}, \text{Low} >$. For now, we can also assume that agents learn from in-group members only. I.e., a yellow will only copy the strategies of other yellows.[14]

What happens when we evolve this game? The first important observation is that the stable end points, or evolutionary equilibria, are different from those described in the single population model in section 2. Now we have equilibria that specify, 1) strategies within the first group, 2) strategies within the second group, and 3) between-group strategies. Within each group, the stable equilibria mimic those for a single population. The greens, for example, might all make fair demands of each other, or settle on the fractious equilibrium. And ditto the yellows. This is because within-group evolution just recreates the conditions of a single population. Between groups, there are three stable equilibria, one where both groups make fair demands of the other, one where the yellows demand High and greens Low, and one where the yellows demand Low and the greens High.

These latter two equilibria can be thought of as bare bones representations of a discriminatory convention or norm. The groups make stable demands that systematically advantage one group at the cost of the other. Once they have reached these equilibria, no individuals in the model can unilaterally change behaviors and improve their outcome. If those receiving Low try to make Medium demands, for example, they get the disagreement point. This tracks the sorts of conditions that oppressed groups often find themselves in vis norms for dividing resources (Cudd, 1994; Okin, 1989).

A number of papers have confirmed that when groups evolve in these models, they often arrive at these discriminatory conventions. Axtell et al. (2001) develop an early model of this sort. Their actors evolve via a kind of *best response dynamic* where they randomly meet partners for bargaining, remember what strategies those partners used, and then in each interaction try the strat-

[14]This type of in-group learning is not actually a necessary assumption for the main relevant results. For example, we will consider models from Axtell et al. (2001) where actors learn individually to repeat actions that benefit them, and where inequitable conventions emerge.

egy that would do best against this interactive history. In their model, two groups most often evolve to the fair equilibrium, but also consistently evolve to the two unfair equilibria. They contrast this with single population models, and point out that otherwise irrelevant tags groups can stabilize unfairness and discrimination. O'Connor (2019), in her book *The Origins of Unfairness*, makes this point thoroughly, exploring a number of different models in the process. As she points out, under many reasonable conditions, this basic model will evolve to unfair outcomes with high probability. A number of other models, focused on a variety of related phenomena, confirm the general picture developed here. When groups divide resources, the presence of social categories means that inequity can emerge endogenously (Bowles and Naidu, 2006; Stewart, 2010; Poza et al., 2011; Hwang et al., 2014; Cochran and O'Connor, 2019; Heydari Fard, 2022).

Another set of relevant results focuses not on the Nash demand game, but on breaking symmetry in social coordination problems. In some situations, actors need to use complementary strategies in a game to do well, but one strategy is preferable. A classic case is the hawk-dove game pictured in figure 2a. This game assumes that opponents can take an aggressive strategy (hawk) or a passive one (dove). The two equilibria, bolded in this figure, are Hawk vs Dove and Dove vs Hawk. At either of these equilibria it is better to be a hawk. Hoffmann (2006) and Amadae and Watts (2022) both illustrate how groups with categories playing this game can evolve to situations where one side always plays dove, to their disadvantage. We see the same in simple complementary coordination games, like the one pictured in 2b. In this scenario, two actors must take complementary strategies, A and B, to succeed. This might represent division of labor, where A involves one set of jobs and B a complementary set. A population with two groups, say men and women, might evolve to solve this problem when one group always plays A (engages in market labor) and the other B (focuses on household labor). But when one outcome is preferable, say $B > A$, this leads to persistent advantage for one group (O'Connor, 2019). In a group without categories, these outcomes are not possible. Thus categories allow for coordination on a new set of efficient equilibria. But they also allow for categorical inequity that would not otherwise be possible.

Taken together, the results described here show that critiques from theorists like Mills and Pateman are supported by models of cultural evolution. Whenever social categories like gender and race are in place, fairness is not what we typically expect from natural social contracts. Insetad, we often expect unfairness. Given that social categories are culturally ubiquitous we thus should not expect emergent social contracts to be fair. To clarify, the claim here is not that these models exactly match the pictures presented by either author. Rather, they confirm a picture where cultural divisions that ought not impact resource distribution in an ideal, just social contract tend to nonetheless become deeply relevant to natural, emergent contracts.

Mills, in particular, is not describing a "racial contract" between white people and oppressed races (as the models here might represent). Instead, his racial contract is among white people, and excludes all others. He supports this picture

		Player 2	
		Dove	**Hawk**
Player 1	**Dove**	1,1	**1,3**
	Hawk	**3,1**	0,0

(a) The hawk-dove game.

		Player 2	
		A	**B**
Player 1	**A**	0,0	**A,B**
	B	**B,A**	0,0

(b) A simple complementary coordination game.

Figure 2: Two games where social categories can underpin inequitable norms. The (pure strategy) Nash equilibria are bolded.

by drawing on the history of colonialism and white domination, where members of non-white countries were often brutally oppressed, enslaved, or killed. These harms were often perpetrated by countries that espoused classic liberal values, though the relevant rights were clearly not extended to colonial subjects. The models to this point, where individuals freely interact to bargain, and learn from the results of their interactions, do not capture the dynamics of oppression at play in this history.

In the rest of this section, we will consider a number of extensions and variations on this model. These will illustrate how unfair contracts emerge under different conditions, and their ubiquity. In the next section, we look at situations that better match Mills' history—one group has power over the other, and this power shapes emergent bargaining outcomes.

3.1 Power

As noted, one thing missing from the models discussed to this point is the coercive nature of the way inequitable contracts are often formed in reality. There is no coercion in these models, and there is no sense of power inequity between groups. Part of what makes them such effective epistemic tools, in fact, is the way that otherwise entirely identical groups starting from neutral states ("of nature") can evolve to stable, discriminatory norms. But we still might wish to know: what happens if we add power to these models?[15]

Power is a complex concept, and has been extensively analyzed by philosophers. A typical distinction is drawn between power-to and power-over, where

[15]Power is a central topic in social contract theory. For example, see Lloyd in this volume, who looks at incorporating realistic power relations into the new social contract theory.

the former tracks the personal capabilities of an agent, and the latter the ability of one agent to force or coerce another to their will (Allen, 2022). Here we will consider a typical way to build power into evolutionary bargaining models. As I will outline, it is consistent with both interpretations of power.

Power is most often included in bargaining models via disagreement points. (These, remember are the payoffs agents receive when their demands are jointly too aggressive.) Agents with different disagreement points will end up in different positions should bargaining fail, which impacts how much the bargain matters to them, and which, in turn, impacts their power over the bargain. This goes back to Nash (1953), who interprets these differences as related to power-over. On his interpretation, before bargaining, each agent makes some threat about what they will do if bargaining fails. More powerful threats translate into lower disagreement points (or *threat-points*) for the other agent. Thus agents can use coercion (or power-over) to reshape their opponent's bargaining position.[16] Alternatively, agents might have different disagreement points because of material or political differences in their lives that make bargains more or less important to them. In highly influential work on household division of labor, for example, Manser and Brown (1980) and McElroy and Horney (1981) represent women as having lower disagreement points in the case of divorce because of economic and legal disempowerment. Thus disagreement points can also be interpreted as a difference in power-to.

Figure 3 shows the Nash demand game with different disagreement points for players 1 and 2. Bruner and O'Connor (2018) build an evolutionary model where two groups evolve to bargain, one with a higher disagreement point ($D > d$). They find that this power imbalance systematically advantages the more powerful group, who tend to end up at the outcome where they demand High more often. The greater the power, the greater the discrepancy. This happens because powerful individuals have relatively little incentive to adopt low demands—their disagreement point is not much worse. As a result they move towards such demands more slowly, and tend to end up adopting higher demands instead.

LaCroix and O'Connor (2020) use agent-based models to show how this can happen even when power is unevenly distributed among the groups. Imagine a situation where just a few men use violence to gain household bargaining advantages. Or imagine that some members of a racial group are economically advantaged, even though the rest are not. They show that in these cases an entire group can, nonetheless, end up with a bargaining advantage. This happened when individuals bargaining with powerful actors extend what they learn to others in the same social category.

Note that, unlike Nash's analysis, power is not guaranteed to lead to bargaining advantage in these models. The powerful group still will sometimes end up getting less due to accidents of learning history. O'Connor (2019), though, shows how power can compound into persistent advantage for one group. She

[16] Nash (1953) uses a rational-choice based approach to argue that power of this sort will lead to better bargaining outcomes for a player. He specifies a set of axioms that govern each player and derives a bargaining solution on the basis of them. On this solution, a player with a higher disagreement points gets more resource.

		Player 2		
		Low	**Med**	**High**
Player 1	**Low**	3,3	3,5	3,7
	Med	5,3	5,5	D,d
	High	7,3	D,d	D,d

Figure 3: Payoff matrix for a simplified Nash demand game where player 1 has a higher disagreement point than player 2, and is thus more powerful.

develops models where a group that gains a bargaining advantage today has a higher disagreement point tomorrow (due to economic empowerment), and is thus more likely to gain further bargaining advantages. In her model, one group always ends up persistently advantaged over the other. And initial power imbalances are highly likely to translate into persistent ones.[17]

Note that the impacts of power in this sort of model do not capture many of the oppressive dynamics described by Mills (1997). But they do help illustrate how power differentials can translate into bargaining asymmetries in natural contracts. A general take-away is that power differences between social contracts increase the chances that natural social contracts will be unfair.

3.2 Minority Status

Bruner (2017) shows how minority status can also impact the likelihood that one group ends up with more than another in emerging bargaining conventions. In his models, two groups evolve to play the Nash demand game, but one is smaller than the other. He finds that in many scenarios, the large group ends up discriminating against the small one. This is the result of learning speed differences between the two groups. Minority members meet their out-group more frequently than majority members do, and thus learn to interact with them more quickly. In bargaining scenarios, this often translates into learning safe, accommodating demands, which the majority group can then learn to take advantage of. This has been called the *cultural red king effect* in reference to an analogous effect in models of natural selection (Bergstrom and Lachmann, 2003).[18]

Depending on details of the game, sometimes a small group can actually gain an advantage via the same effect. This will happen when it tends to yield better payoffs, on average, to make risky, high demands. In these cases, the small group moves more quickly towards demanding High, and the larger group slowly learns to accommodate. O'Connor (2017) points out, though, that the cultural red king effect will generally tend to exacerbate existing inequalities. Disadvantaged groups tend to be risk averse, since they are less economically stable. This tends to lead to minority disadvantage in evolved bargaining, because small, risk averse groups quickly move towards accommodating demands. In addition,

[17] For more on the ways that power can compound and reproduce itself see Tilly (1998).
[18] See also O'Connor and Bruner (2017).

if both groups have tendencies towards in-group favoritism and out-group bias, being in a smaller group tends to lead to general bargaining disadvantage. (See also Amadae and Watts (2022).)[19]

O'Connor et al. (2019) show how minority effects and power effects can interact and create special sorts of disadvantage for some groups in natural contracts. In their models, agents have intersectional identities in that they are part of more than one social category (gender and race being a key example). Following previous theorists, they explore the possibility that at the intersections of disadvantage, i.e., when someone is part of both a minority group and a disempowered group, special levels of disadvantage may arise (Collins and Chepp, 2013; Collins and Bilge, 2020). They find, indeed, that occupying an intersectionally disadvantaged group in these models significantly decreases the bargaining payoffs one expects to get.

Both Young (1993b) and Gallo (2014) find a related effect as a result of information asymmetries between groups. Young (1993b) finds that if one group has more memories than the other, they tend to switch strategies less often and, as a result, to benefit from a version of the cultural red king effect. Gallo (2014) finds the same when one group in better networked than the other. In situations where one group has special sources of information—"old boys clubs" or similar social networking tools—they may gain an advantage in evolving bargaining conventions.

There is a general observation to pull out at this point. Power asymmetries, size asymmetries, and informational asymmetries in these models all increase the chances that one group will end up discriminating against the other. In doing so, they all decrease the chances that fair conventions emerge. Since the real world tends to be full of asymmetries, and very short on symmetries, we can take the findings across these sections to support a picture where unfairness is the expected outcome of natural contracts.

3.3 Emerging Categories

So far, the models in this paper have assumed 1) that social categories are already in place, 2) that social categories are binary, 3) that it is easy to recognize and respond to the category membership of others, and 4) that individuals cannot change social categories. There are a few important things to recognize. First, not all social categories work this way. Some are flexible, changeable, or hard to identify. And second, all social categories are shaped by the processes of cultural evolution. Understandings of race, for example, are deeply culturally shaped and vary across cultures. If we go far enough back into our evolutionary history we will find some version of humans with sexually differentiated behavior, but with no genders. Part of the emergence of unfair natural contracts is the emergence of the type of categories that can underpin these contracts.

[19] Mohseni et al. (2021) find in an experimental setting that smaller groups do tend to be disadvantaged by emerging bargaining norms. They use this finding to argue that processes which at every stage adhere to tenets of historical justice a la Nozick (1974) can nonetheless end up looking very unfair. More on this in the conclusion.

Popa et al. (2021) consider variations on these models where the properties of categories can vary. First, they consider conditions where actors may change categories by adopting different tags. In this sort of situation, unfair equilibria are not possible. Individuals getting low payoffs will switch to a group that gets higher payoffs, re-establishing a one-population model. Central to the establishment of unfair norms is the requirement that a disadvantaged group is stuck with their identity. Popa et al. (2021) show that, in particular, when bargaining happens between more powerful and less powerful groups the powerful are incentivized to adopt tags that differentiate themselves in order to stabilize their bargaining advantages. Popa et al. (2021) also look at a model where agents cannot always correctly identify category membership. They show that the more ambiguous category membership is, the less likely unfairness is to evolve.[20]

Bright et al. (2022) expand these two sets of results to consider the ways that powerful groups might be incentivized to shape or choose categories for oppression, focusing on race. Their models show that powerful groups will tend to learn to pay attention to category markers that are unambiguous and hard to change. They argue that this helps explain why race is a locus of oppression in capitalist systems, in line with the theory of racial capitalism (Cox, 2001; Táíwò et al., 2021). Race, as developed in these systems, is fairly recognizable and fairly hard to change, and this stabilizes inequity. Thus part of the reason racial categories exist, and take the forms they do, relates to the emergence of oppressive systems.

O'Connor (2019) is interested not the in emergence of racial categories, but gender categories. She uses two-population models to show how gender can emerge in most societies to solve coordination problems, such as those related to division of labor. She points out that in a one population model actors who adopt gender tags, and condition their behavior based on these tags, can outperform others. Thus gender spreads. Saunders (2022a) points out that her models assume gendered learning, i.e., that (as in many of the models we have considered so far) agents copy strategies from those in their own social categories. But, he asks, how might this sort of learning have emerged before it was useful to divide labor? He develops agent-based models that show how gendered learning can co-evolve with gendered behavior.[21]

Unfair natural contracts depend on the existence of social categories, but social categories themselves are the products of cultural evolution. The literature just described shows us that such categories can emerge themselves, sometimes in concert with natural contracts.

[20] See also Bruner (2017).

[21] Saunders (2022b) explores more generally the conditions under which agents will come to learn from in-group versus out-group members. He shows that this sort of learning is most stable (and most useful) when groups play "anti-coordination" games, like those that represent division of labor. Notably, bargaining games also have an anti-coordination character in that for all but the fair equilibrium, actors must coordinate on different strategies. This may point to a picture where inequal norms, social categories, and in-group learning stabilize each other.

4 Modeling Social Change

Most social contract theorists believe that justified social contracts should be fair. And many people espouse norms of fairness and claim these norms ought to hold across societies.[22] Given this, we might ask: Does the framework described in this paper tell us anything about social change? Does it tell us how to form fairer natural contracts?

The models indicate that without categorical divisions, unfair contracts are unlikely to emerge. If we were to remove categorical divisions from societies, then, perhaps we would destabilize unfair contracts and promote fair ones? This sort of proposal is typically impractical. In addition, there are a number of arenas where recognizing social categories is crucial for reforms aimed at addressing inequity. But the work described above from Popa et al. (2021) and Bright et al. (2022) suggests ways that categories might be weakened to a point that they do not effectively support unfair norms. Where possible, social rules for categorization that allow the adoption of new tags and markers should weaken systems of inequity. The same is true for social rules that create more ambiguity about category membership.[23] This sort of reimagining of categories is a promising avenue for social change.

Another important take-away from these models is that systems of inequity are *equilibria*, and in this way they are stable and self-reinforcing. As noted, authors like Okin (1989) and Cudd (1994) emphasize that oppressed women often cooperate with their oppression because that is the rational thing to do. In the unfair equilibria modeled here, a disadvantaged individual gets their highest possible payoff by demanding Low of their out-group. Their social environment prevents them from doing well by demanding High (or even making fair demands). This means that in such systems minor changes meant to improve equity—like moral education, or anti-bias training—may not be particularly effective. Instead, for change to happen a significant number of individuals must actually take different sorts of actions. They must demand more (or less) and thus create social responses that push towards a different pattern of behavior (O'Connor, 2019). This may help explain why social change so often involves protest, revolution, and action-based resistance. Interventions to facilitate such actions, such as creating places where marginalized groups can communicate and coordinate, may be effective (Bright et al., 2022).

Another important take-away comes from models of asymmetries, and especially power asymmetries. These models suggest that unfair outcomes are more likely to emerge, and are more stable, when groups are more asymmetric.

[22] See Yaari and Bar-Hillel (1984). In addition, in experimental settings subjects playing the Nash demand game pick fair demands almost exclusively, even in cases where they might gain advantage by demanding more (Nydegger and Owen, 1974; Roth and Malouf, 1979; Van Huyck et al., 1997).

[23] One might respond that when it comes to, say, race, this will not always be possible or desirable. This is true. On the other hand cultural rules like "sumptuary laws" were intended to prevent oppressed racial groups from adopting the dress and styles of privileged groups, and thus creating more racial ambiguity (Pastore, 2002). There are lots of ways that race can be more or less ambiguous without disappearing (and ditto gender).

Thus, we might expect that actions intended to promote equity that do not disrupt underlying power dynamics are unlikely to be effective. Laws aimed at wealth redistribution, and legal reforms to protect the democratic influence of marginalized groups, should accompany calls for ethical reform if such reform is to be successful. The models indicate that reparations will be more effective than implicit bias training. Bright (2022) critiques those who spill too much ink talking about racism without advocating for actual material changes for racially oppressed groups. The models here suggest his critiques are on the right track.

A last important take-away, emphasized by O'Connor (2019), is that inequity emerges robustly across a wide range of models under very minimal preconditions. These preconditions are that 1) individuals recognize social categories, 2) they condition their behavior on social categories, and 3) they learn to take actions that benefit themselves. Since all these features are likely to be present in most social groups, we should expect that underlying social dynamics will tend to persistently push towards inequity. Thus, attempts to eradicate inequity are unlikely to be permanently successful. We should thus adopt a model where unfairness is something to be continually watching for, and continually combating, rather than something that will someday be "fixed". Inequity is a hydra whose heads grow back.

5 Conclusion

Natural contracts emerge when people in a society learn and culturally evolve to solve social problems. We see natural contracts related to cooperation, joint action, division of labor, and communication. A key type of natural contract regards division of resources—how will groups make decisions about who gets how much? While previous work has shown that fair contracts tend to emerge on culturally evolutionary time-scales, this paper argues that the reverse is true in groups with social categories. When there are loci that a group can use to coordinate inequity, they will tend to learn to do so.

Inequity of this sort emerges endogenously, and under common conditions. Furthermore, various asymmetries between groups, especially related to power, make inequity more likely, and more stable. The take-away is that natural contracts will tend to be unfair. As discussed in section 4, this framework for understanding unfair natural contracts has implications for how we might think about establishing fairer ones.

The work discussed here is mostly relevant to descriptive work on the emergence of actual natural contracts. But it can also speak to work on justifying contracts. Nozick (1974) gives an account of justice that is based in history. A distribution of wealth derived from just processes will be a just one. Although Nozick is not very specific on what processes count as just, arguably the models here track such a process.[24] Actors freely make requests for resources, and are not forced into bargaining, or changing their bargaining demands. However,

[24]He specifies that they will involve just initial acquisition of resources, and just transfers of holdings.

the outcomes that emerge advantage some social groups at the cost of others, for no justificatory reason. This advantage emerges as an accident of history, as a result of symmetry breaking in an evolutionary process. Such outcomes push against historical accounts of justice by showing how arguably just processes produce social outcomes that we think of as inequitable and undesireable (Mohseni et al., 2021).[25]

In general, the work described here also supports a methodological point. Natural contracts often emerge over long timescales, and as the result of many, distributed interactions in a large population. As such, they emerge from processes that are highly complex and hard to study. Armchair reasoning about these processes is not likely to be informative (as we see with Nozick (1974)). Evolutionary models can help by providing concrete tools to study how natural contracts might emerge, and how we should understand them.

Acknowledgements

Many thanks to participants at the New Social Contract Theory workshop for comments on this work. Special thanks to John Thrasher and Michael Moehler for comments and their work putting this volume together.

References

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

Alexander, Jason (2007). *The Structural Evolution of Morality*. Cambridge University Press.

Alexander, Jason and Brian Skyrms (1999). "Bargaining with neighbors: Is justice contagious?." *The Journal of philosophy*, *96*(11), 588–598.

Allen, Amy (2022). "Feminist Perspectives on Power." *The Stanford Encyclopedia of Philosophy*. Ed. Edward N. Zalta and Uri Nodelman. Fall 2022 edition. Metaphysics Research Lab, Stanford University.

Amadae, SM and Christopher J Watts (2022). "Red Queen and Red King Effects in cultural agent-based modeling: Hawk Dove Binary and Systemic Discrimination." *The Journal of Mathematical Sociology*, 1–28.

Arrow, Kenneth J (1978). "Nozick's entitlement theory of justice." *Philosophia*, *7*(2), 265–279.

[25]This sort of worry was recognized as early as Arrow (1978) who wrote, "Suppose a dominant group, say whites or "Aryans", agreed to trade with the complementary minority only on very unfavorable terms. Indeed, they might not have to agree in any concrete sense: suppose each one happened for his own reasons to resolve to so act...Are we to say that the results are just?" (272).

Axtell, Robert, Joshua M Epstein, and H Peyton Young (2001). "The emergence of classes in a multi-agent bargaining model." *Social Dynamics*, 191–211.

Bergstrom, Carl T and Michael Lachmann (2003). "The Red King effect: when the slowest runner wins the coevolutionary race." *Proceedings of the National Academy of Sciences*, *100*(2), 593–598.

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

Binmore, Kenneth George (1994). *Game theory and the social contract*. MIT press.

Bowles, Samuel and Suresh Naidu. Persistent institutions. Technical report, working paper, Santa Fe Institute, (2006).

Bright, Liam Kofi (2022). "White Psychodrama.".

Bright, Liam Kofi, Nathan Gabriel, Cailin O'Connor, and Olufemi Taiwo (2022). "On the Stability of Racial Capitalism." *Ergo*.

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

Bruner, Justin P (2017). "Minority (dis)advantage in population games." *Synthese*, doi 10.1007/s11229–017–1487–8.

Bruner, Justin P and Cailin O'Connor (2018). "Power, Bargaining, and Collaboration." *Scientific Collaboration and Collective Knowledge*. Ed. Conor Mayo-Wilson Thomas Boyer and Michael Weisberg. Oxford University Press.

Carvalho, Jean-Paul and Bary Pradelski (2019). "Identity and underrepresentation: Interactions between race and gender." *Available at SSRN 3299477*.

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

Collins, Patricia H. and Valerie Chepp (2013). "Intersectionality." *The Oxford Handbook of Gender and Politics*. Ed. Johanna Kantola Georgina Waylen, Karen Celis and S. Laurel Weldon. Oxford University Press, chapter 2, 57–87.

Collins, Patricia Hill and Sirma Bilge (2020). *Intersectionality*. John Wiley & Sons.

Cox, Oliver Cromwell (2001). "Caste, class and race: A Study in social dynamics." *Racism: Essential Readings*, *7*, 49.

Cudd, Ann E (1994). "Oppression by choice." *Journal of Social Philosophy*, *25*, 22–44.

D'Arms, Justin, Robert Batterman, and Krzyzstof Górny (1998). "Game theoretic explanations and the evolution of justice." *Philosophy of Science*, 76–102.

Gallo, Edoardo. Communication networks in markets. Technical report, Faculty of Economics, University of Cambridge, (2014).

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

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

Hwang, Sung-Ha, Suresh Naidu, and Samuel Bowles. Social conflict and the evolution of unequal conventions. Technical report, Working Paper, (2014).

LaCroix, Travis and Cailin O'Connor (2020). "Power by association." *Ergo*.

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

McElroy, Marjorie B and Mary Jean Horney (1981). "Nash-bargained household decisions: Toward a generalization of the theory of demand." *International economic review*, 333–349.

Mills, Charles W (1997). "The racial contract." *The Racial Contract*. . Cornell University Press.

Mohseni, Aydin, Cailin O'Connor, and Hannah Rubin (2021). "On the emergence of minority disadvantage: testing the cultural Red King hypothesis." *Synthese*, *198*(6), 5599–5621.

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

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

Nozick, Robert (1974). *Anarchy, state, and utopia*. Volume 5038. new york: Basic Books.

Nydegger, Rudy V and Guillermo Owen (1974). "Two-person bargaining: An experimental test of the Nash axioms." *International Journal of game theory*, *3*(4), 239–249.

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

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

O'Connor, Cailin and Justin Bruner (2017). "Dynamics and Diversity in Epistemic Communities." *Erkenntnis*, (doi: 10.1007/s10670-017-9950-y).

Okin, Susan Moller (1989). "Justice, gender, and the family." *Justice, Politics, and the Family*. . Routledge, 63–87.

O'Connor, Cailin, Liam Kofi Bright, and Justin P Bruner (2019). "The emergence of intersectional disadvantage." *Social Epistemology*, *33*(1), 23–41.

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

Pateman, Carole (1988). "Sexual contract." *The wiley blackwell encyclopedia of gender and sexuality studies*, 1–3.

Popa, Mihaela, Roland Muhlenbernd, and Jeremy L. Wyatt (2021). "Fairness and Signaling in Bargaining Games.".

Poza, David J, José I Santos, José M Galán, and Adolfo López-Paredes (2011). "Mesoscopic effects in an agent-based bargaining model in regular lattices." *PloS one*, *6*(3), e17661.

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

Roth, Alvin E and Michael W Malouf (1979). "Game-theoretic models and the role of information in bargaining.." *Psychological review*, *86*(6), 574.

Saunders, Daniel (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, Daniel (2022b). "When is Similarity-biased Social Learning Adaptively Advantageous?.".

Skyrms, Brian (1994). "Sex and justice." *The Journal of philosophy*, 305–320.

Skyrms, Brian (1996). *Evolution of the social contract*. Cambridge University Press.

Skyrms, Brian (2004). *The stag hunt and the evolution of social structure*. Cambridge University Press.

Stewart, Quincy Thomas (2010). "Big bad racists, subtle prejudice and minority victims: An agent-based analysis of the dynamics of racial inequality." *Annual Meeting of the Population Association of America*. .

Sugden, Robert et al. (1986). *The economics of rights, co-operation and welfare*. Springer.

Táíwò, Olúfémi O, Anne E Fehrenbacher, and Alexis Cooke (2021). "Material Insecurity, Racial Capitalism, and Public Health." *Hastings Center Report*.

Tilly, Charles (1998). "Durable inequality." *Durable Inequality*. . University of California Press.

Van Huyck, John B, Raymond C Battalio, and Frederick W Rankin (1997). "On the origin of convention: Evidence from coordination games." *The Economic Journal*, *107*(442), 576–596.

Yaari, Menahem E and Maya Bar-Hillel (1984). "On dividing justly." *Social choice and welfare*, *1*(1), 1–24.

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

Young, H Peyton (1993b). "An evolutionary model of bargaining." *Journal of Economic Theory*, *59*(1), 145–168.