Philosophical Studies
https://doi.org/10.1007/s11098-023-02096-1

Taught rules: Instruction and the evolution of norms

Camilo Martinez[1]

Accepted: 19 December 2023

Abstract
Why do we have social norms—of fairness, cooperation, trust, property, or gender? Modern-day Humeans, as I call them, believe these norms are best accounted for in cultural evolutionary terms, as adaptive solutions to recurrent problems of social interaction. In this paper, I discuss a challenge to this "Humean Program." Social norms involve widespread behaviors, but also distinctive psychological attitudes and dispositions. According to the challenge, Humean accounts of norms leave their psychological side unexplained. They explain, say, why we share equally, but not why we disapprove of those who don't. I defend the Humean Program against this challenge. In particular, I suggest an idea for how to extend the Program to account for the psychological side of norms. Socially adaptive behaviors aren't just likely to emerge in a group; They are also likely to be widely taught within it. The transmission of these behaviors through instruction explains why they are associated with distinctive normative attitudes and dispositions. These attitudes play a pedagogical role in helping transmit these behaviors to children and newcomers.

Keywords Social norms · Cultural evolution · Evolutionary models · Normativity · Instruction

1 Introduction

Having agreed to go hunting together, the Lion, the Fox, the Jackal, and the Wolf manage to catch a stag. If each of them contributed equally to the hunt, how should they divide up the kill?

For most of us, this isn't a hard question: Into four equal parts, of course! That's what *fairness* requires. Indeed, absent more information about the case, any other division seems unfair. Thus, we resent the Lion when we learn he's taken the lion's share of the kill.

✉ Camilo Martinez
camilom@princeton.edu

[1] Department of Philosophy, Princeton University, Princeton, NJ, USA

Published online: 27 January 2024

A harder question is why this answer comes so naturally to us. Why do we tend to think that, in this situation and others like it, one *should* share alike? From where comes the idea that doing otherwise is *wrong*?

According to many philosophers, these ideas are best accounted for in *evolutionary* terms. We disapprove of unequal division because we have somehow evolved to think so. But the kind of evolution at stake here isn't biological. It is *cultural* evolution, which means, roughly, a change in belief over time (Alexander, 2007, 19).

I call the project of explaining our commitment to fairness and other social norms in cultural evolutionary terms "The Humean Program." The name comes from David Hume's account of the origin of the rules of justice, particularly the institution of property, in Book 3 of *A Treatise of Human Nature*. There, Hume argues that these rules aren't the outcome of an explicit agreement between rational parties. Instead, they arise out of a gradual process of social evolution. As he writes:

> Nor is the rule concerning the stability of possession the less deriv'd from human conventions, that it arises gradually, and acquires force by a slow progression, and by our repeated experience of the inconveniences of transgressing it. (2007, sect. 3.2.10)

Modern-day Humeans apply this insight to explain the origin of various social norms. Besides norms of fairness, they have studied norms of cooperation and coordination, trust, property, and gender.[1] In each case, Humeans aim to offer a naturalistically plausible account of how these norms first emerged.

Yet, despite its promise, the Humean Program faces a significant challenge. As some critics have argued, the program incurs an *explanatory deficit*. Humeans aim to explain our commitment to social norms by showing that certain behaviors associated with them are highly likely to emerge and stabilize in a group through cultural evolution. However, explaining why people tend to behave in certain ways isn't yet to explain why they endorse a *norm* to the effect that one should act that way. It doesn't explain, for example, why they consider deviations from such behaviors to be *wrong*, not just unexpected. Insofar as social norms are more than mere behavioral propensities, the Humean Program must be found lacking.

In this paper, I defend the Humean Program against this challenge. In particular, I put forward a promising idea for how to extend the Program to account for the "normativity" of social norms. To be clear, I don't offer an actual extension. As we will see, doing so will involve substantial theoretical and empirical work.

Put briefly, the idea is this: Socially adaptive behaviors like equal sharing aren't just likely to evolve in a group. They are also likely to be widely *taught* within it. Once they emerge, group members are likely to transmit these behaviors to newcomers by monitoring and correcting how they act. But this type of instruction profoundly transforms the nature of these practices. Monitoring and correction imply the type of psychological attitudes and dispositions in virtue of which a behavioral

[1] For fairness, see Skyrms (2014, chap. 1), Alexander (2007, chap. 5), Zollman (2008), and Binmore (2005). For cooperation and coordination, see Axelrod (1984), Alexander (2007, chap. 3), Young (1993), and Sugden (2004). For trust, see Skyrms (2014, chap. 3) and Alexander (2007, chap. 4). For property, see Gintis (2007). For gender, see O'Connor (2019).

regularity counts as a social norm proper. Given instruction, practices like equal sharing become shared *rules*, not just common behaviors.

The structure of the paper is this. In Sects. 2 and 3, I describe the Humean Program, focusing on one specific application of it: Brian Skyrms' influential account of the origin of our norm of fair division. Next, in Sects. 4 and 5, I raise the explanatory challenge and argue against some ways Humeans have responded to it. Finally, in Sects. 6–10, I develop my idea for how to meet the challenge, focusing again on the case of fairness.

2 The Humean Program

The guiding hypothesis of the Humean Program is that social norms are culturally evolved solutions to recurrent problems of social interaction (Alexander, 2007, 23; Skyrms & Zollman, 2010, 266). Humeans aim to explain the presence of a social norm in a group in terms of that norm's cultural success in promoting socially adaptive behaviors. Confirming this hypothesis for any given social norm N present in group G involves establishing three different claims.

First, Humeans must establish *Recurrence*. They must show that there is a type of social situation, S, such that members of G regularly find themselves in it. In other words, S is a recurrent social situation in the group.[2] To describe the central features of S, Humeans usually model it as a *game* in the sense of game theory. A game in this sense refers to a set of two or more agents or "players," each of whom can adopt one of several different action strategies. In a game, each player's outcome or "payoff" depends on the combination of strategies adopted by all the players. Humeans model different types of social situations using different games.

Second, Humeans must establish *Nuisance*. They must show that whenever members of G find themselves in S, they face a social interaction problem, p. In other words, p is a problem inherent to the social situation. Usually, p involves some tension between individual rationality and collective action. It is a situation where each person acting rationally doesn't ensure that they act together successfully. In every such case, group members must find some way of coordinating or cooperating.[3,4]

Finally, Humeans must establish *Resilience*. They must show that norm N has culturally evolved among the members of G because it allows them to respond to problem p in situation S. In other words, N is a culturally adaptive response to the social interaction problem. To show this, Humeans usually offer an *evolutionary model* of the population of G when they are repeatedly faced with cases of S. At the

[2] S might refer to a specific type of social situation or a broad *class* of social situations. Here, I focus on the former case for simplicity's sake. However, the latter approach promises to make sense of why there is often ambiguity and conflict in the application of social norms. See Skyrms and Zollman (2010, 266) for discussion.

[3] In other cases, S may be problematic not because it involves a tension between individual rationality and collective action but because group members lack the required information or willpower to implement the uniquely rational course of action for everyone. I thank a reviewer for helping me see this point.

[4] For an accessible introduction to some familiar games used to model problematic social situations, see O'Connor (2022, sect. 4).

minimum, such a model must include a representation of the state of the population at any given time, as well as a specification of the dynamical laws describing how that population state changes over time (Alexander, 2007, 25). With such a model in hand, Humeans aim to show that norm *N*—represented in the model as a particular action strategy—is *evolutionarily robust*, meaning that it is very likely to emerge and stabilize in the population given plausible assumptions. Ideally, one wants to show that *N* emerges as a stable evolutionary equilibrium in the model under a maximally wide range of initial conditions.[5]

Humeans use different types of evolutionary models.[6] Although some of them admit a biological interpretation, they are meant to be models of cultural evolution. This term means, roughly, a change in cultural traits (i.e., beliefs, behaviors, skills) over time (Alexander, 2007,19). Cultural evolution is possible because some cultural traits are more adaptive than others, allowing people to cope better with social life. Furthermore, adaptive cultural traits are transmitted more often between group members, thus spreading in the population. The transmission of cultural traits relies on different forms of *social learning*, namely, learning where people acquire beliefs or behaviors from others (Laland & Hoppitt, 2013; Henrich & McElreath, 2003). Cultural traits spread in a group because people acquire them from others through mechanisms such as imitating those who are more successful than them.

For Humeans, establishing Recurrence, Nuisance, and Resilience goes a long way toward explaining a social norm's presence in a group. Taken together, these three claims offer an *equilibrium explanation* of the norm (Sober, 1983). This type of explanation is both *less* and *more* informative than a standard causal explanation, where an event (like the emergence of a norm in a group) is explained by citing its cause. Unlike a causal explanation, an equilibrium explanation doesn't specify the causal history of the explanandum. Yet, it makes up for this lack of detail by showing that a specific causal history isn't needed to account for the explanandum's presence in a given setting. Insofar as the explanandum is a robustly emergent equilibrium in that setting, it is highly likely to be brought about by some cause or another. In other words, equilibrium explanations present a *disjunction* of causal scenarios (1983, 204). Thus, they are more informative than explanations that focus on a single causal history.[7]

[5] My notion of an evolutionarily robust norm or strategy is similar to Sober's (1983) notion of a *global equilibrium* and Alexander's (2000) notion of a *stochastically robust strategy*.

[6] See Alexander (2007, chap. 2) for an overview.

[7] Even if Humeans models fall short of explaining how a social norm potentially evolved, in the sense of specifying the conditions under which it likely emerged and stabilized in a group, they might still offer other types of valuable information. For instance, they may offer insight into how the norm possibly evolved or about the minimal conditions required for it to evolve. For this distinction between "how-potentially," "how-possibly," and "how-minimally" uses of evolutionary models, see O'Connor (2019, sect. 0.2).

3 Skyrms on fairness

Moving forward, it will be useful to see how one can apply the Humean Program to a specific case. I'll focus on Brian Skyrms' (2014; 1996) account of a widespread norm of fair division according to which people with equal claims to some good should receive equal shares of it. Skyrms' account of this norm is one of the best developed and most discussed applications of the Humean Program. As such, it will be my focus for the remainder of the paper.

Skyrms' account of fair division starts with Recurrence. He describes a type of social situation that people in any human group are likely to encounter. These are cases where two or more individuals must decide how to divide a good between them on pain of losing it. Moreover, the parties to this interaction are all symmetrically positioned in that all have a similar level of strength, experience, speed, etc. (Kitcher, 1999, 223). Imagine, for instance, a group of hunters who, having caught a stag together, must decide how to divide it between them before it goes to waste.[8]

Skyrms introduces a version of the *Nash bargaining game* as a formal model of this simple distribution problem. This game involves two players who must figure out how to divide a chocolate cake between them. They must do so by each independently writing a final claim to a percentage of the cake on a piece of paper and handing it to a referee. If both claims add up to more than 100%, the referee eats the cake. Otherwise, each player gets what they claimed.

Moving to Nuisance, Skyrms notes that players in this divide-the-cake game face an *equilibrium selection problem*. Without communicating, they must somehow agree on a way of dividing up the cake. If they ask for more than 100%, they get nothing. If they ask for less, they fail to get as much cake as they can. The problem is that there are many ways for them to jointly ask for *exactly* 100% of the cake. They could both ask for half, or one could ask for 60% and the other for 40%, etc. Each such combination of strategies is a *Nash equilibrium* of the game, in the sense that each player does as well as they can do given what their partner is doing. The players' problem consists in coordinating to select one among many such equilibria.

Turning to Resilience, Skyrms wants to show that in a group regularly facing the divide-the-cake game the strategy of demanding half of the cake is highly likely to evolve. He argues for this claim by offering an evolutionary model. In particular, he models the population according to the *replicator dynamics*. This model assumes that different game strategies (Demand Half, Demand 60%, etc.) are initially present in the population with different frequencies. Moreover, the frequency of any given strategy increases or decreases depending on how the expected fitness of those who follow it compares to the average fitness of the population (Alexander, 2007, 28). Here, one's expected fitness is measured in terms of one's expected payoff in the

[8] For more contemporary examples of situations with this structure, see Alexander (2007, 150).

game. So, if a strategy is such that its players can expect to do better in the game than the group average, then the proportion of people playing it increases accordingly.[9]

The replicator dynamics can be interpreted in strictly biological terms (Taylor & Jonker, 1978) However, the model also admits a cultural interpretation (Weibull, 1995). Indeed, one can show that when a group uses a particular social learning strategy, their cultural evolution approximates the replicator dynamics (Schlag, 1998). This learning strategy consists in imitating others with a probability proportional to their success in the game. Someone who follows this learning rule randomly selects another player from the population and then compares her own payoff to theirs. If the other person's payoff is higher, the player adopts their strategy with a probability proportional to the payoff difference (Alexander, 2021).

Having introduced this model, Skyrms shows that demanding half of the cake is an evolutionarily robust strategy in the population under the assumption that strategies in the divide-the-cake game are weakly and positively correlated. According to this assumption, people playing the same strategy are slightly more likely to partner up with each other than with people who follow a different strategy.[10] There are several ways of defending this assumption in Skyrms' model. Social interactions may happen within the same family or clan or in social networks involving "neighborhoods" of like-minded individuals (Alexander, 2000). In any case, under these conditions, the strategy of demanding 50% of the cake emerges as a stable evolutionary equilibrium under virtually all the initial conditions of the model. Thus, Skyrms concludes that:

> In a finite population,[11] in a finite time, where there is some random element in evolution, some reasonable amount of divisibility of the good and some correlation, we can say that it is likely that something close to share and share alike should evolve in dividing-the-cake situations. This is, perhaps, a beginning of an explanation of the origin of our concept of justice. (1996, 21)

[9] More formally, the replicator dynamics represents the state of the population at any given time using a state vector $\vec{s} = (s_1, \ldots, s_m)$, where s_i denotes the proportion of group members who follow strategy i. The expected fitness of an agent following i ($F(i|\vec{s})$) is their expected payoff, namely, the sum of the payoffs they would obtain playing every other strategy in the population, weighted by their probabilities. The crucial dynamical assumption of the model is that the instantaneous rate of change of s_i is a function of the difference between the expected fitness of agents who play that strategy and the average fitness of the population. In mathematical form:

$$\frac{\partial s_i}{\partial t} = s_i(F(i|\vec{s}) - F(\vec{s}|\vec{s}))$$

In this equation,

$$F(\vec{s}|\vec{s}) = \sum_{i=1}^{m} s_i F(i|\vec{s}).$$

For how to derive this equation, see Alexander (2007, 29).

[10] In Skyrms' model, correlation works through a function that inflates the likelihood that a strategy meets itself and deflates the likelihood that it meets a different strategy. For details, see Skyrms (1996, chap. 1, fn. 30).

[11] Strictly speaking, Skyrms' model posits an infinite population, so this summary of his view seems to be inaccurate in this respect. This point is made by Alexander (2007, 159).

4 The explanatory deficit

Humean accounts of the origin of social norms can fail in several ways. First, regarding Recurrence and Nuisance, any such account might fail to be *representative* (D'Arms, Batterman, & Gorny, 1998, 89). It might posit a situation *S* and a problem *p* that the group is unlikely to have ever encountered. Hence, norm *N* can't be explained as an evolved response to situations with that structure. For example, some critics argue that Skyrms' divide-the-cake game is an unrealistic model of distribution problems. Although these problems are frequent in social life, they seldom involve a referee or near-perfect symmetry between the parties (Kitcher, 1999, 223; D'Arms, Batterman, & Gorny, 1998, 89-90).[12]

Further, regarding Resilience, a Humean model might fail to be sufficiently *robust* (1998, 90). The model might fail to show that norm *N* is likely to emerge and stabilize in the population across a sufficiently wide range of conditions. For instance, D'Arms, Batterman, and Gorny (1998) argue that Skyrms' model of fair division is insufficiently robust. As we saw, this model assumes that action strategies in the divide-the-cake game are weakly correlated. Yet, people might rationally want to play this game in a way that introduces anti-correlation between their behaviors. Someone who demands 60% of the cake might seek a partner who asks for 40%, not one who asks for 60%. Assuming anti-correlation, Demand Half is much less likely to evolve in the replicator dynamics.

Even if Humean accounts are representative and robust, they face a deeper problem. Indeed, this problem affects the Humean Program in general, not this or that specific model. As some critics have pointed out, the program seems to incur an *explanatory deficit*. Humeans aim to explain our commitment to social norms (e.g., our norm of fair division) by showing that certain action strategies associated with them (e.g., sharing equally) are evolutionarily robust. However, explaining why people tend to act in a certain way isn't yet to explain why they endorse a *norm* to the effect that one should act that way. Social norms aren't mere behavioral regularities. Hence, fully accounting for them requires more than explaining the behaviors associated with them.[13]

Social norms go beyond mere behavioral regularities in at least two ways. First, when a social norm exists in a group, people adopt a special attitude towards deviations from the regularity in behavior. They consider such deviations to be *wrong*, not just unexpected (Hart, 2012; Anderson, 2000). Second, because people think of deviations this way, they're disposed to punish or otherwise sanction those who deviate (D'Arms, 2000). Consider our norm of fair division. When someone takes more than their fair share of a good—like the Lion in Aesop's fable—we're not only surprised by their behavior. We also think they have acted wrongly. Moreover, we usually resent them and may even call them out on their behavior. In sum, social norms have a *psychological* dimension. They are behavioral regularities undergirded by distinctive attitudes and dispositions.

[12] For a response to these criticisms of Skyrms' model, see Alexander (2000, 150).

[13] For different versions of this critique, see Kitcher (1999), D'Arms (2000), and Anderson (2000).

Humean models of the evolution of norms leave their psychological side unexplained. These models address the evolution of behavior, not psychology (O'Connor, 2022). That is, while they show that certain behaviors are likely to culturally spread in a population on account of being socially adaptive, they're silent about the psychological mechanisms behind them. As far as the models go, the behaviors might be caused in a variety of ways. A behavior like sharing equally *could* be produced by the type of attitudes and dispositions characteristic of social norms, but it could also be caused by something else. It might, for instance, result from a "fast and frugal," System 1 heuristic, like the rules of thumb commonly used by chess players (Alexander, 2007, 22-23).[14] This insensitivity to the psychology behind culturally evolved behaviors impacts the explanatory power of Humean models. As Justin D'Arms puts the point, "a model that is avowedly insensitive to whether the behavior it explains is (regarded as) [normatively] significant cannot be an explanation of that significance" (2000, 299).

One may wonder whether Humeans really incur an explanatory deficit. As we saw, Humean models of the origin of norms are meant to be models of *cultural* evolution. But these models represent interactions between cultural agents, that is, agents whose behavior is guided by normative beliefs, values, and emotional tendencies. Given this assumption about the kind of agents they apply to, it seems we can interpret Humean models as describing the evolution of behaviors to which agents attach certain normative significance from the outset. In short, we can interpret them as models of the evolution of norms proper, not mere regularities of behavior.

However, even if we can attribute normative attitudes and dispositions to the agents who figure in Humean models, the question is whether these models accord any role to such attitudes in the emergence of the relevant practices. The answer would seem to be "no" (D'Arms, 2000). In Humean models, cultural fitness is *exclusively* a function of how agents behave in social interaction problems. Hence, whatever selection pressures these models target, they act primarily on behaviors. Due to these pressures, some behaviors proliferate, others die out, irrespective of whether group members have normative attitudes towards them or not. What guarantees, then, that they will have such attitudes?[15]

Maybe this criticism of the Humean Program is too quick. Perhaps Humean accounts of norms can be extended to account for their psychological dimension. As we saw, Skyrms only claims to offer "a *beginning* of an explanation of the origin of our concept of justice" (1996, 21; emphasis added). This claim suggests that he thinks more work is needed to give a complete account of how fairness evolves.

However, the problem with the Humean Program isn't just that it incurs an explanatory deficit but that it makes it hard to see how this deficit could be bridged. As we saw, Humeans aim to show that strategies like equal sharing are evolutionarily robust. But this means that these behaviors must be highly *stable* in the face of deviations. In Skyrms' model, for example, any "mutant" or innovator who decided

[14] Some Humeans explicitly compare social norms with such heuristics. See Alexander (2007, 22-23).

[15] As critics of the Humean Program point out, this problem need not arise for all evolutionary models of the origin of norms. Some models may accord a relevant role to normative attitudes in the emergence and stabilization of socially adaptive behaviors. See, e.g., Gibbard (1982).

to adopt a strategy other than Demand Half would be driven into extinction pretty quickly. The reason for this is simple: Once Demand Half has taken over the group, playing a different strategy simply doesn't pay. Anyone who deviates from the regularity is guaranteed to have a lower expected payoff than the population average.[16]

Under these circumstances, it is hard to see what would be the *point* of the type of attitudes and dispositions characteristic of social norms (D'Arms, 2000). If cultural evolution already weeds out deviations from the regularity in behavior, why would group members ever come to disapprove of such deviations and be disposed to sanction those who deviate? Having such attitudes and dispositions would be *costly* for them. Blame and disapproval are psychologically taxing (Shoemaker & Vargas, 2021). Imposing sanctions on others takes time and effort (Elster, 1989; Buchanan, 1975). Yet, these attitudes and dispositions would seem to play no useful role in the group. They are not needed to sustain the practice.

One could argue that normative attitudes and dispositions play a role in the evolutionary dynamics described by Humean models. According to this idea, if playing a strategy different from Demand Half doesn't pay in a group, then this is partly because one will suffer other people's disapproval and sanctions. So, plausibly, part of the reason why cultural evolution weeds out deviations from fairness has to do with these attitudes and dispositions.

However, this just isn't so. In Skyrms' model, strategies other than Demand Half go extinct due to the replicator dynamics. Yet, this dynamics makes absolutely no reference to disapproval or sanctions, only to the agent's expected payoff in the divide-the-cake game. If a strategy isn't copied as much as others, this isn't because people who play it get punished more often, but because they get less cake on average. As far as the model goes, disapproval and sanctions are irrelevant to the cultural evolution of fairness.

In sum, the Humean Program faces a dual challenge. It incurs an explanatory deficit because it fails to explain the psychological dimension of social norms. Moreover, it makes it hard to see how this deficit could be bridged because it seems to leave no space for normative attitudes and dispositions to play a social role or function.

[16] Indeed, in the divide-the-cake game, Demand Half is the only *evolutionarily stable strategy* in the sense of Maynard Smith and Price (1973). Roughly, a strategy is evolutionarily stable just in case it is able to withstand invasion by mutants once it has taken over a population. This doesn't mean, however, that Demand Half is the only evolutionarily stable state in the game. As Skyrms shows, there are "polymorphic" states involving more than one strategy that are evolutionary stable. Part of his account of the evolution of fairness focuses on specifying the conditions under which a population can avoid such "polymorphic traps." For a generalization of the idea of an evolutionarily stable strategy, see the notion of an *evolutionarily stable set*, i.e., a set of strategies that remains stable in a population even if there is some drift between the different strategies in the set (Thomas 1984).

5 Sympathy and resentment

Humeans haven't failed to notice this challenge to their project. Indeed, Hume himself was aware of the need to account for the attitudes of approval and disapproval behind social norms. In the *Treatise*, he distinguishes between two questions we can ask about the origin of justice. First, how do the rules (viz., regularities) of justice get established among us? And second, why do we come to "attribute to the observance or neglect of these rules a moral beauty and deformity"? (2007, sect. 3.2.1). In other words, why do we come to think of acts of justice and injustice in terms of *right* and *wrong*? As we saw, under the Humean Program this second question arises for social norms in general. Humeans have proposed different answers to it.

Hume's own answer appeals to the emotion of *sympathy*. We disapprove of deviations from regularities like equal sharing because we feel for the people who are harmed by them. Once these regularities exist in a group, deviating from them hurts other people's interests. In a population where Demand Half is the statistical norm, a "greedy" mutant who asks for 60% of the cake will disadvantage most people who interact with him. The mutant asks for more cake than is customary, thus causing others to get no cake at all. In Hume's view, group members condemn greedy behavior out of sympathy for those negatively affected by it.

In the latest edition of *Evolution of the Social Contract*, Skyrms suggests a similar explanation for why the practice of equal sharing is likely to become a norm. He writes: "If the equal split is a convention in dividing-the-cake situations, it is no surprise that greedy players should be despised or ostracized, since they spoil things for those with whom they interact" (2014, 22).

However, this account doesn't quite work. In divide-the-cake games where one party acts fairly and the other acts greedily, *both* end up empty-handed. They jointly ask for more than 100% of the cake. Yet, we're only inclined to feel sympathy for the fair one. The reason for this, I take it, is that we think that she's the one who's been wronged. In other words, sympathy for the fair party presupposes a belief that her interests aren't just frustrated, but *wrongfully* so. Otherwise, we would feel sympathy for the greedy party, too. But then people who sympathize with others in this type of case must already think that it is wrong to be greedy.

Robert Sugden (1998) proposes a similar view. When we have good reason to expect that others will satisfy our preferences, we tend to *resent* them if they don't. Sugden calls these expectations that trigger resentment *normative expectations*. They shouldn't be confused, he says, with beliefs to the effect that someone should do something.[17] For Sugden, normative expectations are just reasonable beliefs that someone will do something where one prefers that they do that thing. Practices like equal sharing give rise to such expectations. When equal sharing takes over a population, people can reasonably expect that others will partake in this behavior. Moreover, they prefer that they do so because it makes it easier to coordinate with them. Hence, they will resent those who don't share equally because they act against their normative expectations.

[17] Nor with beliefs that others believe that one should do something, as the term is used by Bicchieri (2006).

Sugden's view is similar to David Lewis's (1969) account of how coordinating conventions can turn into norms. Lewis writes:

> ...if [other people] see me fail to conform [to a convention], not only have I gone against their expectations; they will probably be in a position to infer that I have knowingly acted contrary to my own preferences, and contrary to their preferences and their reasonable expectations. They will be surprised, and they will tend to explain my conduct discreditably. The poor opinions they form of me, and their reproaches, punishment and distrust are the unfavorable responses I have evoked by my failure to conform to the convention. (1969, 99)

This view falls prey to the same problem as the sympathy view. Suppose that in previous years you've always bought me a fancy present for my birthday. Accordingly, I now believe that you'll do the same this year. If you act against my expectation, I may feel frustrated and displeased. Yet, I wouldn't resent you for it unless you also *promised* to get me a fancy gift. In short, to trigger disapproval, expectations require a belief that others are under some obligation to satisfy them. But this means that people who resent others in the divide-the-cake game must already think that they should share equally.

The problem with these two views is that they try to account for the attitudes and dispositions associated with social norms in terms of emotions and emotional responses that presuppose their existence. Hence, they fail to meet the first part of the challenge to the Humean Program. They offer an explanation for the psychological side of norms, but it is viciously circular.

Further, these views fail to meet the second part of the challenge, viz., that of elucidating the role or function of normative attitudes and dispositions. The views depart from the spirit of the Humean Program. The guiding hypothesis of the program, recall, is that social norms are culturally evolved solutions to social problems. According to these accounts, however, disapproval and sanctions don't arise in a group in response to a problem. Instead, they are additions to an independently evolved solution. The idea is that once a behavior like equal sharing evolves, it interacts with people's emotional tendencies in a way that leads them to regard it as normatively significant. But this interaction isn't guided by the type of selection pressures that explain the initial emergence of the behavior. The normativity of the practice "evolves" in the sense of developing gradually, but it isn't *selected for*.

Is there a way for Humeans to discharge their explanatory burden? Can the Humean Program be extended to account for the type of normative attitudes and dispositions behind social norms? In the remainder of the paper, I suggest a possible extension. The key to the extension, I believe, lies in the relationship between cultural evolution and social learning.

As we saw above, cultural evolution is possible because human beings are social learners: We can acquire beliefs and behaviors from each other. Indeed, according to various theorists, our capacities for social learning evolved through natural selection precisely because of the role they play in cultural evolution. These capacities allow us to accumulate a body of adaptive knowledge across generations—a *culture*. And having access to this knowledge increases our fitness (Henrich & McElreath, 2003).

For example, it gives us access to skills and information that no individual could acquire on her own.

The crucial insight behind the Humean Program is that we should think of norms as *cultural products*—as a part of this body of adaptive knowledge. Norms encode *socially* adaptive information, i.e., information allowing us to cope better with social life. As such, they come to be transmitted between people through social learning. Eventually, they become part of our common lore.

My idea for how to extend the Humean Program has to do with the specific way norms are socially transmitted. As we saw, Humean models like Skyrms' tend to assume that social norms are transmitted through imitation. People learn to share equally because they copy other people who share equally. I suggest that social norms are also likely to be transmitted through *instruction*. People learn to share equally, not just because they copy others, but because those others *teach* them how to play fair.

The transmission of behaviors like equal sharing through instruction profoundly transforms their nature. Or so, at least, I wish to suggest. Teaching someone how to play fair involves monitoring and correcting how she acts. But these activities imply normative attitudes and dispositions. These attitudes and dispositions play a *pedagogical role* in helping transmit the practice.[18]

In the next few sections, I sketch what my suggested extension to the Humean Program might look like, focusing on the case of our norm of fair division. My goal isn't to work out all the details. Instead, I want to offer reasons to think that the details *can* be worked out. You may take what follows as a proof of concept.

6 Population replacement

In Humean fashion, I'll start with a version of Recurrence. I'll describe a type of situation that human groups are likely to encounter once they have evolved practices like equal sharing. As I'll argue, groups are prone to facing forces that destabilize the evolutionary equilibria behind these practices

Consider one way Humean models fail to be representative of real human groups. These models usually abstract away from the phenomenon of *population replacement*, namely, the process whereby people in a group are replaced by others over time. Population replacement can happen for several reasons. For example, it can result from *migration*, with some people immigrating to and others emigrating from the group. Or it can happen due to *generational change*, with newer generations gradually succeeding older ones.

Plausibly, most real human groups experience some degree of population replacement over time. People often move in and out of neighborhoods. Families and clans

[18] Other theorists have already suggested that there is a link between instruction and normativity. For example, in his genealogical account of norms, Philip Pettit (2023, chap. 1) claims that coordinating conventions may become rules in virtue of people using these practices to regulate each other's behavior. Likewise, teaching plays a prominent role in Jonathan Birch's (2021) "skill hypothesis" concerning the evolution of normative cognition. Finally, Castro and Toro (2014) argue that teaching is crucial for the emergence of cumulative cultural evolution.

frequently gain and lose members. Yet, Humean models of the evolution of social norms don't usually take these processes into account. For instance, in Skyrms' model of the origin of fairness, group practices change over time because group members update their strategy depending on their expected payoff, not because anyone gets replaced by a different player.

Some Humean models involve a mutation parameter, μ, representing the probability that any given group member switches strategy after each round of play. In principle, this parameter can be used to represent the random substitution of group members by others who play different strategies (Kandori, Mailath, & Rob, 1993). In practice, however, the parameter is mostly used to model processes like innovation or experimentation with new strategies. Hence, it is assigned a small value. In J. McKenzie Alexander's (2000) agent-based model[19] of the evolution of fairness, for instance, the value of μ is as low as 0.001. This means that in a population of 1,000 people, only about one person will change strategy after each round of the divide-the-cake game. We should expect the effect of population replacement to be greater than this. Migration and generational replacement seem like more significant sources of behavioral variation within a group than innovation or experimentation.[20]

Population replacement can lead to the introduction and elimination of game strategies, thus changing their relative frequency in a group. It should be clear how emigration and deaths lead to the elimination of strategies from a group. But how can immigration and new births introduce game strategies? On the one hand, new arrivals may come from groups that haven't (yet) evolved the relevant practices. Or, upon migrating, they might switch strategies thinking that circumstances in the new group warrant a change. Someone who used to behave fairly might turn greedy because they hope to get more cake than they used to. Or they might switch to being modest, asking for only 40% of the cake, wanting to curry favor with their new peers. Children, on the other hand, are unfamiliar with the type of social situation the practice solves. Hence, they may approach it using strategies that make sense in other circumstances. Faced with the divide-the-cake game, they might use the otherwise sensible strategy of trying to get as much of the good as possible.

By eliminating and introducing game strategies, processes like migration and generational change can affect the stability of evolutionary equilibria. As we saw, according to Humeans, evolutionarily robust strategies like equal sharing are highly likely to emerge and stabilize in a group. Yet, migration and generational replacement can act as external forces that cause perturbations in such equilibria. A population might reach an equilibrium only to be displaced away from it by the inflow

[19] Unlike the replicator dynamics, which models a population using an aggregate state vector, agent-based models represent a population discretely. They include information about each individual group member, such as what strategy they use and how they are spatially or socially positioned. See Alexander (2007, chap. 2) for discussion of the different types of evolutionary models.

[20] Of course, if we want to use μ to model population replacement, we cannot estimate an adequate value for it simply from the armchair. We need a better understanding of the phenomenon. For example, we need to estimate how often population replacement happens relative to how often group members face the relevant social interaction problem. This already points to an observation I'll make below about how extending the Humean Program in the direction I suggest will require substantial theoretical and empirical work.

and outflow of group members. Being an evolutionarily robust state, the equilibrium is likely to withstand such perturbations. But it might be subject to further perturbations. In the long run, a practice can undergo cycles of stabilization and destabilization.[21]

As an illustration of this phenomenon, consider Figures 1 and 2, both of which show the evolution of a population of 1,000 agents who play 200 rounds of Skyrms' divide-the-cake game.[22] For simplicity, only three strategies are represented in the population: The *Fair* strategy (Demand Half), the *Modest* strategy (Demand 40%), and the *Greedy* strategy (Demand 60%). Initially, every strategy has the same frequency. In each round of this model, agents randomly play the game with others and then update their strategy using the imitative learning rule described in Sect. 2, namely, they imitate others who have a higher payoff than them with a probability proportional to the difference between their payoffs.

Figure 1 shows how the population evolves without much population replacement ($\mu = 0.001$).[23] As we see, under these conditions, fairness quickly emerges and stabilizes in the group, following Skyrms' results. The population *converges* to fairness in the sense of Alexander (2000), that is, in the long run, everyone except $N \times \mu$ group members behaves fairly. Figure 2, in contrast, shows how the population behaves under conditions of population replacement. More precisely, 10% of the population is replaced by others with random strategies after each "generation" or round of play ($\mu = 0.1$). Under these conditions, the population never converges to fairness: The proportion of people who behave fairly is always below 90%. Moreover, the population seems to go through cycles of stabilization and destabilization: It starts to converge to the equilibrium only to be displaced away from it later.

[21] The way population replacement affects the stability of a practice may be influenced by the specific type of social interaction problem the practice evolved to solve. For example, situations with the structure of a Stag Hunt may be destabilized differently than situations with the structure of a Prisoners' Dilemma. I thank a reviewer for bringing this point to my attention.

[22] These models were created using Abed-1pop, an agent-based modeling framework developed by Izquierdo, Izquierdo, & Sandholm (2019). You can download Abed-1pop here: https://luis-r-izquierdo.github.io/abed-1pop/. Please email me at camilom@princeton.edu for a file with the specific parameters I used in my models.

[23] N.B. In these models, I use parameter μ to represent different rates of population replacement instead of mutation. I follow Kandori, Mailath, & Rob's (1993) suggestion that this parameter can be used to model the random substitution of group members with others who play different strategies, such as immigrants. Of course, random replacement is an idealizing assumption. Very plausibly, new group members don't update their strategies in a purely random way. For example, children's strategies are likely influenced by their parents. More sophisticated models may represent population replacement differently. They may, for instance, explicitly represent the flow of individuals between different interacting populations. Here, I only want to illustrate the effect that population replacement might have over a practice like equal sharing, so I abstract away from these complications. More realistic models of the phenomenon must take these details into account.

7 Imitation and myopia

I'll now move to a version of Nuisance. I'll argue that, given frequent population replacement, groups with practices like equal sharing may face a problem concerning how they readapt to these practices after they have been destabilized. This problem has its roots in the social learning strategy group members use to acquire behaviors from each other.

In the divide-the-cake game, Demand Half is the *optimal* strategy for everyone to adopt in the long run (Alexander, 2000). Given the evolutionary dynamics at play, everyone is better off playing fair. Under other states, fair players to get less than 50% of the cake because they are likely to meet a few greedy players. Modest players can't expect to get more than 40%, come what may. And, as the population approaches fairness, the payoff of greedy players goes well below 50%. Hence, when a population is displaced away from the state of widespread fairness, it's in everyone's best interest to go back to that state as quickly as possible.

But a group may not readapt to fairness as quickly as desirable. The reason comes down to the type of social learning rule people use to update their strategy. As we saw, in Humean models people are usually assumed to use some sort of imitation rule to learn from others. One common rule of this sort consists in imitating others who are more successful than them with a probability proportional to the difference between their payoffs. Yet, this kind of rule can be *myopic* in the following sense: It can lead some of the population to shift towards a strategy whose fitness benefits are only transient. This is because a strategy may have an advantage over others simply because of current peculiarities of the population, not because it is the best strategy in the long run (Alexander, 2021). Still, for a while, the strategy gets copied, thus spreading in the group.

Given population replacement, imitative social learning can lead some group members to shift to modesty, even if they'll eventually return to fairness. This process might work as follows.

At first, population replacement introduces new greedy players to the group. This increase in the frequency of greedy players gives modesty a slight advantage over fairness. Modest players do better against greedy players than fair players do. Playing against a greedy player, a modest player gets 40% of the cake. In contrast, a fair player gets no cake. In the short term, some of the population moves towards modesty. For example, if I'm a fair player who happens to encounter a greedy player, I'll fail to get any cake. If I then compare my payoff to that of a modest player, the imitative learning rule will lead me to switch from fairness to modesty.

Eventually, however, greedy players go extinct. Although they do well against modest players, they do very poorly against fair players and other greedy players. Once this happens, fairness becomes the best strategy to play in the game. No matter who they play against, fair players get half of the cake while modest players only get 40% of it. So, the population shifts back to fairness.

In brief, once a population is displaced away from fairness due to population replacement, its road back to it can have a "detour" through modesty. Some group members may temporarily switch to modesty because this strategy does better

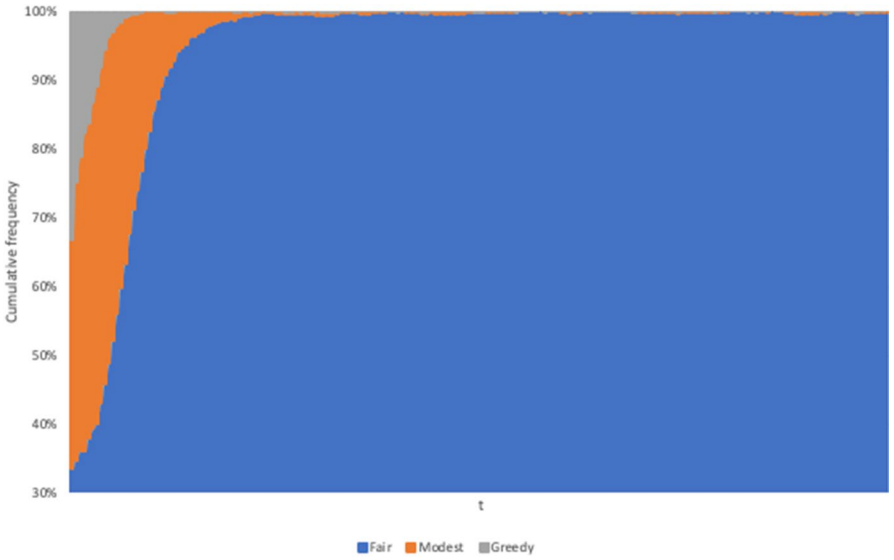

Fig. 1 Evolution of fairness without population replacement (N=1000; μ=0.001; t=200)

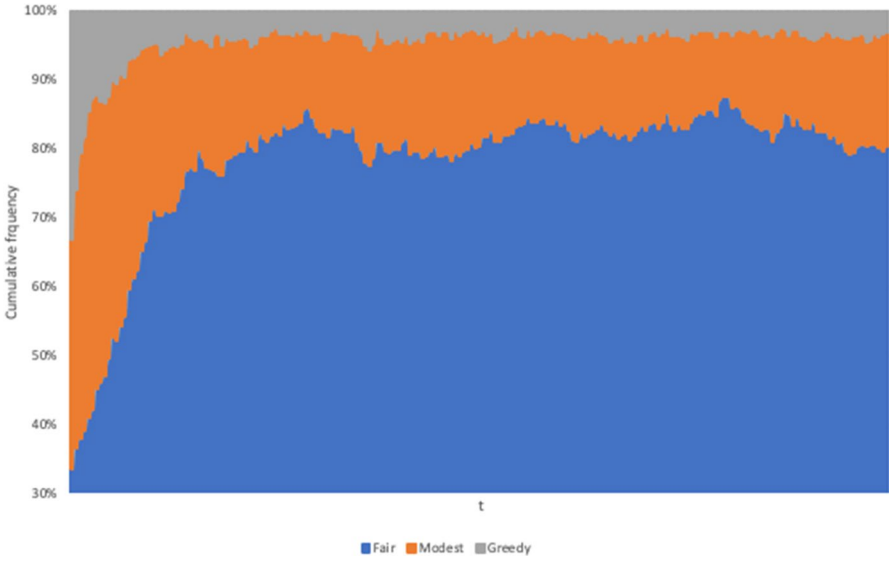

Fig. 2 Evolution of fairness with population replacement (N=1000; μ=0.1; t=200)

against greediness. But this detour gives rise to a kind of social problem. Because of it, the population readapts to fairness more slowly than it otherwise could, to everyone's disadvantage.[24]

8 Instruction and natural selection

Finally, I turn to a version of Resilience. I'll argue that faced with the problem described in the last section, human groups may develop a way of responding to it, namely, they may come to use a different social learning strategy. Instead of acquiring behaviors like sharing equally merely through imitation, human populations may evolve a practice of acquiring them through instruction.

Human groups may respond in several ways to forces that threaten to unravel socially beneficial practices like equal sharing. Indeed, a large literature in the social sciences and game theory studies various mechanisms that communities might use to stabilize cooperation and other socially beneficial arrangements. For example, in response to the proliferation of non-cooperators in a group, group members may respond by directly withholding cooperation from defectors (Axelrod, 1984; Axelrod & Hamilton, 1981; Trivers, 1971), keeping track of people's reputations (Nowak & Sigmund, 2005; Leimar & Hammerstein, 2001; Boyd & Richerson, 1989), and developing community enforcement practices, such as various forms of punishment and exclusion (McAdams, 1997; Hirshleifer & Rasmusen, 1989; Fudenberg & Maskin, 1986).

Of course, a group may respond in all these ways to the sort of destabilization of its practices brought about by population replacement. I believe, however, that the problem described in the last section creates a sustained pressure for groups to develop a new and distinct solution. This problem, recall, specifically had to do with how behaviors are socially transmitted in a group. In particular, the problem arises due to the myopic character of social learning through imitation. Given the recurrent nature of this problem, we might expect groups to develop a solution that addresses it at its root by somehow improving how behaviors are socially transmitted. This solution, I want to suggest, is instruction.

As discussed in Sect. 2, Humean accounts of the origin of norms rely on the idea that people are social learners: They have a capacity for acquiring beliefs and behaviors from each other. As we saw, one important type of social learning is imitation. In it, people acquire a behavior by copying the actions of another person, the "model." This type of behavior transmission doesn't require the model to do anything to facilitate imitation. Indeed, she need not be aware that she's being emulated. But social learning can also happen through *instruction* (Sterelny, 2012; Tomasello, Kruger, & Ratner, 1993). Here, the model of the behavior takes steps to facilitate its

[24] Imitation can also give rise to other problems. For example, the learning rule of imitating the best may favor adopting the strategy of the best agent in one's local neighborhood, who may turn out to have a very poor strategy from the global point of view. (I thank a reviewer for bringing this point to my attention). Imitation may also be problematic if individuals are prone to making mistakes when copying the actions of others (Castro and Toro 2014) As I discuss in Sect. 10, one question that arises under my proposed extension to the Humean Program is exactly what sort of limitations of imitative social learning give rise to a need for instruction.

transmission to others. She may do things like demonstrating the behavior to them, giving them instructions on how to perform it, and correcting them if they get it wrong.

Under what conditions should we expect a behavior to be taught, rather than merely copied? On the one hand, instruction has various benefits relative to imitation. First, it can increase the fidelity of behavior transmission. With imitation, there's always a risk that one might get what the model is doing wrong. This risk decreases, however, if that person is involved in the learning process. Moreover, instruction can pass along behaviors that are impossible to acquire otherwise. For example, "negative" behaviors, such as refraining from picking one's nose, can't be copied. One needs to be taught not to engage in them. Finally—and crucially for our purposes—instruction can transmit behaviors more effectively than imitation. Many behaviors are acquired more easily and quickly if they're taught rather than copied. Think, for instance, of how hard it would've been for you to learn how to tie your shoes without anyone's assistance.

On the other hand, instruction also has various costs. In particular, it is costly for the model, who must take actions aimed at handing down the practice. Indeed, instruction is a form of cooperation, in that one individual engages in a costly behavior that benefits others with no immediate benefit to herself (Thornton & Raihani, 2008, 1825).

One natural idea is this: Instruction is favored by selection pressures only if the long-term benefits teachers can expect to obtain from teaching outweigh the costs to them of doing so (2008,1826). The expected benefits of teaching depend on the difference that teaching makes to how effectively pupils acquire a behavior. In particular, instruction has a high utility only if it increases the likelihood and ease with which pupils acquire the behavior relative to other forms of learning. For example, adult meerkats obtain long-term fitness benefits in terms of kin selection if pups learn how to handle prey effectively. Moreover, pups would have a very hard time picking up this skill through imitation. Hence, in meerkat populations, natural selection likely favors the transmission of prey handling through a form of teaching where adult meerkats let pups handle some of the prey they catch (2008, 1827).

Now, consider culturally evolved practices like equal sharing. Once these practices evolve, population replacement seems to create a pressure for them to be passed down through instruction, not just imitation. First, group members obtain fitness benefits if newcomers to the group learn how to share equally, especially if they pick up this behavior quickly. In particular, if new arrivals and children learn how to play fair soon after joining the group, then everyone avoids the "detour" through modesty in the road back to fairness. The practice of fairness is destabilized by the inflow of new people, but the group quickly readapts to the optimal equilibrium. Moreover, new group members are unlikely to pick up this behavior quickly enough through imitation. Indeed, imitation is what causes the detour due to its myopic nature. Under these conditions, the benefits group members can expect to obtain from instructing newcomers in the practice may well outweigh the costs they incur in doing so.

But we need to tread carefully here. Upon reflection, there's an important disanalogy between instruction as it happens in meerkats and how a practice like equal

sharing might be taught among humans. In meerkats, natural selection favors the teaching of prey handling because adult meerkats who instruct their pups in this skill have higher expected fitness than those who don't (plausibly because adults with competent children have a higher expected number of grandchildren). Among humans, however, instructing others in practices like equal sharing seems to be a *public good*. If I teach you how to play fair, then whatever benefits this brings me will also be shared by other group members. If so, then it's hard to see how natural selection might favor a disposition to teach such behaviors. Teachers incur a cost, but this cost doesn't give them a comparative advantage over others.

However, in human beings, natural selection may favor instruction through the process of *cultural group selection*. The idea that cultural groups are subject to group selection pressures is emphasized by several recent theorists of cultural evolution (Richerson et al., 2016; Henrich, 2004; Henrich & McElreath, 2003).[25] According to this idea, groups with certain cultural traits are more likely to grow, resolve internal conflicts, overcome resource constraints, win wars, and replace other groups. Because of these factors, such groups are more likely to survive and flourish than others. Given this type of cultural selection acting on groups, natural selection may slowly favor traits that contribute to group success (Richerson et al., 2016, 5). The presence of these traits makes it more likely that an individual's group will do well. Hence, it indirectly increases their expected fitness. As we just saw, social learning through instruction seems to be one such trait. It contributes to a group's success by allowing its members to respond to the destabilizing effect of population replacement more effectively. If so, then natural selection might favor a capacity for this type of social learning.

To recap, the picture so far is this: The cultural evolution of practices like equal sharing creates a selection pressure favoring a capacity to transmit these practices through instruction. Natural selection favors social learning via instruction because this capacity allows populations to readapt more effectively to socially beneficial equilibria which have been destabilized by forces like migration and generational change.

This general picture fits well with *culture-gene co-evolutionary theory*, one prominent recent approach to explaining the evolution of human sociality (Richerson & Boyd, 2006; Henrich & McElreath, 2003; Feldman & Laland, 1996). According to this theory, many pro-social human traits are best explained as the outcome of two processes working in tandem: cultural and biological evolution. More precisely, the idea is that cultural and biological adaptations co-evolve: Culturally evolved innovations bring about changes in human social environments. These changes give rise to new selection pressures. And these selection pressures shape our minds, making us better equipped for sociality (Kelly & Setman, 2021; Sterelny, 2012). My suggestion

[25] The idea of cultural group selection shouldn't be confused with the idea that natural selection can act directly on groups. This latter idea is highly controversial in biology, with many theorists rejecting it. See Henrich (2004, sect. 4) for discussion. More recently, however, some theorists have attempted to rehabilitate the notion that natural selection may sometimes act on groups. See Sober and Wilson (1998) and Okasha (2006).

is that culturally evolved practices like equal sharing produce new selection pressures favoring social learning through teaching and instruction.

9 The emergence of normativity

But the transmission of equal sharing and similar behaviors through instruction deeply transforms their nature. Instruction brings on stream the type of attitudes and dispositions that distinguish social norms from mere regularities in behavior. As I will argue in this section, instruction involves certain activities on the part of teachers, these activities require them to have certain capacities, and these capacities necessitate normative attitudes and dispositions.

In order to instruct newcomers to a group in a practice like equal sharing, group members must engage in at least two types of activities. First, in their role as teachers, group members must reliably *monitor* the behavior of newcomers. That is, they must keep track of how new arrivals act in divide-the-cake situations, and they must register whenever they deviate from the established practice of fairness. Such monitoring need not require that group members actively seek out transgressors. But it does require that, whenever they witness or participate in interactions with newcomers, they pay increased attention to their behavior and be alert to deviations.

Second, in cases where they detect deviations on the part of newcomers, group members must respond by taking appropriate measures. In particular, they must attempt to *correct* their pupils' behavior, in the sense of intervening to make it less likely that they will deviate again from the practice in the future. Such correction may involve various kinds of interventions. Among other things, group members can indicate that there has been a deviation, impose appropriate penalties, and offer guidance or advice.

The activities of monitoring and correction require group members to have a host of different capacities. On the one hand, monitoring plausibly requires being able to identify who the newcomers are, determine how they behave in the relevant social situations, and keep a record of their past and current behavior. Correction, on the other hand, plausibly requires being able to determine what interventions are appropriate in each particular case, track their effect on pupils, and make adjustments to them as needed.

Besides these general capacities, however, monitoring and correction crucially depend on some other abilities. By focusing on these abilities, we can appreciate how instruction gives rise to normativity.

In particular, monitoring and correction presuppose two types of capacity. Monitoring requires a capacity to make comparisons between *observed* and *expected* behaviors. That is, when engaging in monitoring, group members must be able to compare how newcomers actually behave in divide-the-cake situations with how they are *supposed* to behave in such situations. Otherwise, it wouldn't be possible for them to detect deviations from the practice. Correction, in turn, requires a capacity to modify one's own behavior should the behavior one observes in others be *unexpected*. That is, having detected a deviation, group members must be able to

reorient their behavior towards the transgressor. Otherwise, it wouldn't be possible for them to administer the necessary correctives.

However, to have these capacities, group members must have attitudes and dispositions of a distinctively normative kind. Or so, at least, I wish to argue. First, to make comparisons between observed and expected behaviors, teachers must have *expectations* regarding how people in the group are supposed to behave in a certain class of situations. These expectations can't simply be beliefs about what others are likely to do in the future. They must be *normative* expectations, in the sense of expectations the frustration of which one regards as a kind of failure that calls for an appropriate response.[26] Further, to modify their behavior to respond to such failures, teachers must have various behavioral dispositions. They must be disposed to intervene in relevant ways whenever their pupils don't behave as expected.

As we saw in Sect. 4, when group members have this type of attitudes and dispositions in relation to a practice, this practice isn't a mere regularity of behavior. Instead, group members adopt a special attitude towards it: They effectively treat the practice as a *rule*, that is, as a standard of correct behavior against which the actions of others are to be measured and deviations from which call for some sort of response (Hart, 2012). In teaching newcomers the practice, group members consider deviations from it to be wrong, not just unexpected, and they are prepared to punish or otherwise sanction those who deviate.

This line of argument may raise two worries. First, one could think that teaching a behavior like equal sharing need not involve monitoring and correction. After all, not all forms of teaching involve such activities. Among meerkats, as we saw, adults train pups in how to handle prey simply by letting them play with it. And, among human beings, some forms of teaching rely on "social tolerance" and "opportunity provisioning" (Kline, 2017). Here, teachers don't monitor and correct how pupils act; They simply tolerate them being around and offer them opportunities to learn that they may otherwise not have.

However, these non-corrective forms of teaching might not be sufficient to transmit behaviors like equal sharing in the way required to avoid the detour through modesty described in Sect. 7. Social tolerance and opportunity provisioning would work by offering pupils increased chances to observe and imitate experts in divide-the-cake situations. Recall, however, that in these situations imitation can be myopic, in the sense of leading pupils to acquire behaviors whose fitness benefits are only transient. What is needed is a form of teaching where sub-optimal behaviors are quickly identified and purged. This form of teaching, I think, will likely involve activities of monitoring and correction.

Second, one could argue that, while monitoring and correction do require certain normative attitudes and dispositions, these need not be of the kind that characterizes social norms. These attitudes may be purely *prudential*, in the sense of reflecting ideas about what behaviors are most beneficial or expedient. Teachers may monitor and correct newcomers merely because they think that they must share equally

[26] N.B. These expectations are different from what Sugden (1998), as we saw, calls normative expectations. For him, acting against normative expectations is seen as a failure because it frustrates one's personal preferences, not because it indicates a lack of competence in a practice.

if they hope to do well in the group. They need not think that pupils are under a non-prudential requirement to be fair. In contrast, we think of practices like fair division as involving non-prudential requirements, in the sense that if people don't act fairly, then they're on the hook independently of whether being fair advances their interests.

Note, however, that instruction is likely to be unsuccessful if it relies merely on prudential normativity. As we saw, correction involves imposing sanctions on newcomers when they deviate form the established practice. Yet, prudential normativity is not a good basis for sanctions. Someone who acts against a prudential requirement may always claim that this requirement no longer applies to her because her interests have changed. To be effective, instruction must be associated with a non-prudential form of normativity. Group members must think that not being fair is wrong, but not simply in the sense of being imprudent. They may, for instance, consider it wrong because it goes against local custom: It's at odds with "how we do things around here" (Brennan, Eriksson, Goodin, Southwood, 2013; Pettit, 2019).

In sum, effective social learning through instruction involves the type of attitudes and dispositions characteristic of social norms. These attitudes play a *pedagogical* role in helping us teach practices like equal sharing to others. If natural selection favors a capacity to transmit some behaviors through instruction, then it also likely favors a capacity to hold such attitudes and dispositions. Evolution likely shaped our minds to make us capable of treating some behaviors as rules that people must follow, and be prepared to enforce such rules with sanctions. And this capacity may be the basis for our *norm psychology*, that is, the host of adaptations allowing us to "do norms" by detecting, complying, and enforcing them (Kelly & Setman, 2021).

10 The extended humean program

So far, I have only offered the sketch of an extension to the Humean Program in response to the explanatory deficit challenge. In this section, I'll discuss how the sketch might be completed. The question is: How can my proposed extension be developed into a serious candidate explanation of the normativity of social norms?

Consider my two central claims:

1. Socially adaptive behaviors, such as equal sharing, aren't just likely to emerge in a group through imitation. They are also likely to stabilize in it through instruction.
2. Instruction brings on stream the type of normative attitudes and dispositions that are characteristic of social norms and distinguish them from mere regularities of behavior.

Developing my proposed extension to the Humean Program will require substantiating these two claims. How can this be done? On the one hand, supporting the first claim will require gaining a better understanding of the evolution of teaching. We need to determine under what conditions groups are likely to develop a practice of teaching certain behaviors. And we need to show that these conditions are met in the

case of socially adaptive behaviors like equal sharing. On the other hand, supporting the second claim will require gaining a better understanding of the relationship between instruction and normativity/normative psychology. We need to produce evidence to the effect that these phenomena are, indeed, intimately connected.

The evolution of teaching can be studied through modeling work. Evolutionary models can offer insight into the conditions under which teaching is favored by selective pressures. What sort of models could these be? One approach is suggested by work on the evolution of social learning more generally. Several theorists use evolutionary game theoretic models to study the conditions under which evolution favors social learning over individual learning or innate mechanisms when it comes to acquiring adaptive behaviors (Wakano & Aoki, 2006; Kameda & Nakanishi, 2003; Rogers, 1988). One important lesson from this body of work is that evolution favors social learning when the environment where organisms learn changes at a moderate rate. When the environment changes slowly, social learning isn't worth it: Other learning strategies are equally effective for acquiring adaptive behaviors while being less costly. When the environment changes quickly, social learning isn't useful: There's not enough time for adaptive knowledge to accumulate in the population.

The evolution of teaching could be studied using a similar framework. For example, we might consider models where teaching is compared to imitation across different rates of population replacement. (The idea is that population replacement causes changes in the *social* learning environment individuals face). Intuitively, we should expect the results to be similar to those obtained in social learning models. When the rate of population replacement is low, teaching wouldn't seem to be worth it: The situation is like that of a group facing the odd mutant or innovator who quickly goes extinct. When the rate of population replacement is high, teaching wouldn't seem to be useful—indeed, it may even be harmful: The people doing the teaching are unlikely to possess much adaptive social knowledge.

Evolutionary models can also help us better characterize the problems associated with imitative learning that drive the evolution of instruction. In this paper, I have focused on one such problem: the myopic character of imitative learning rules. But imitation may give rise to other problems, too. Recent work in evolutionary biology studies the conditions under which various problems associated with imitation might create a selective pressure in favor of teaching and instruction. For example, imitation may be problematic when the behaviors that need to be transmitted in a group are complex and difficult to acquire (Fogarty, Strimling, & Laland, 2011) or when pupils are likely to make mistakes in copying the actions of models (Castro and Toro, 2014). My proposed extension to the Humean Program is compatible with this work and may draw from it to better describe the pitfalls of imitation which give rise to a need for teaching as a new, improved mechanism of cultural transmission.

At any rate, once modeling work allows us to spell out the conditions favoring the emergence of teaching, we need to show that these conditions are likely to be met in the case of socially adaptive behaviors like equal sharing. Ideally, we would want to show that, just as such behaviors are likely to emerge across a wide range of conditions, so they are likely to be taught across a similarly wide range of circumstances.

Finally, the relationship between teaching or instruction and normativity might be best explored through empirical research. In particular, my proposal predicts various correlations between our normative tendencies and our pedagogical practices. One of my claims is that normative attitudes and dispositions emerge to serve a pedagogical role. If I'm right, then we should observe correlations between the utility of teaching some behaviors and our tendency to adopt normative attitudes towards them. That is, the normativity of a behavior should be roughly proportional to the utility of teaching it.

This prediction can be tested empirically. For example, the "altruistic punishment" paradigm (Fehr & Gächter, 2002) could be used to determine if people's tendency to punish or impose sanctions on others co-varies with whether these others are likely to modify their ways as a result of being punished. In this experimental paradigm, players in a social dilemma have the choice of punishing others at some cost to themselves. Studies using this paradigm have shown that people tend to punish non-cooperators even when doing so is costly to them and can have no benefits in terms of reciprocity or reputation (that is, the punishment is "altruistic") (see Van Lange, Rockenbach, and Yamagishi, 2014, 6-8). One could test if this tendency varies as a function of whether non-cooperators are "corrigible," in the sense of being capable of learning to cooperate. If my proposal is on the right track, then we should observe less punishment when non-cooperators are (perceived to be) less corrigible.

11 Conclusion

Let me briefly summarize our discussion so far. Modern-day Humeans, as I have called them, believe that social norms, like our norm of fair division, are best accounted for as culturally evolved solutions to recurrent problems of social interaction. This explanatory project, however, faces a challenge. The challenge is showing why culturally evolved solutions to social problems should take a distinctive shape: Why they should be *norms*, as opposed to mere behavioral regularities.

In this paper, I have suggested a way for Humeans to meet this challenge. Norms, as Humeans claim, encode socially adaptive information. This is why they are likely to emerge and stabilize in a group through a form of social learning that relies on imitation. But a group's capacity to access and use this socially adaptive information is compromised by forces like migration and generational change. Hence, there is a pressure for norms to be transmitted through instruction, not only imitation. Normativity emerges in connection to such instruction. Normative attitudes and dispositions play a pedagogical role in the context of activities like guiding and correcting others' behavior.

I believe my suggestion offers a promising way of extending the Humean Program to account for the psychological side of norms. But the suggestion has to be made good. In arguing for versions of their view, Humeans normally advocate for a combination of theoretical and empirical work. In particular, abstract, formal models like Skyrms' replicator dynamics help us formulate the precise conditions under which a practice like equal sharing is likely to emerge and stabilize in a group.

In turn, empirical work helps us determine whether these conditions obtain in the relevant social groups. Likewise, extending the Humean Program in my suggested direction will require a similar combination of theoretical and empirical contributions. We need to develop abstract, formal models that allow us to formulate the precise conditions under which a practice like equal sharing is likely to be transmitted through instruction. And, in addition to this, we need empirical evidence that helps us determine whether these conditions are representative of real human populations and whether instruction is generally correlated with normativity.

In this paper, I have tried to offer reasons for thinking that the extension might be achieved. Whether we can successfully do so, however, remains to be seen. Here, as with many other questions, the proof is in the pudding.

Acknowledgements For helpful discussion, questions, and comments, I'm grateful to Maria Camila Castro, Sam Fullhart, Mark Johnston, Vivian Knopf, Santiago Ospina Celis, Philip Pettit, participants in the Spring 2023 dissertation seminar at Princeton, especially Pietro Cibinel, Rebecca Mullen, Elliot Salinger, Maggie Shea, and Alice van't Hoff, and two anonymous reviewers for this journal.

References

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

Alexander, J. M. (2021). Evolutionary game theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* Stanford University.

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

Anderson, E. (2000). Beyond homo economicus: New developments in theories of social norms. *Philosophy & Public Affairs, 29*(2), 170–200.

Axelrod, R. (1984). *The evolution of cooperation*. Basic Books.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science, 4489*, 1390–1396.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Binmore, K. (2005). *Natural justice*. Oxford University Press.

Birch, J. (2021). Toolmaking and the evolution of normative cognition. *Biology and Philosophy, 36*(1), 1–26.

Boyd, R., & Richerson, P. (1989). The evolution of indirect reciprocity. *Social Networks, 11*(3), 213–236.

Brennan, G., Eriksson, L., Goodin, R. E., & Southwood, N. (2013). *Explaining norms*. Oxford University Press.

Buchanan, J. M. (1975). *The limits of liberty: Between anarchy and Leviathan*. University of Chicago Press.

Castro, L., & Toro, M. A. (2014). Cumulative cultural evolution: The role of teaching. *Journal of Theoretical Biology, 347*, 74–83.

D'Arms, J. (2000). When evolutionary game theory explains morality, what does it explain? *Journal of Consciousness Studies, 7*(1), 296–299.
D'Arms, J., Batterman, R., & Gorny, K. (1998). Game theoretic explanations and the evolution of justice. *Philosophy of Science, 65*(1), 76–102.
Elster, J. (1989). *The cement of society: A study of social order*. Cambridge University Press.
Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415*(6868), 137–140.
Feldman, M. W., & Laland, K. N. (1996). Gene-culture coevolutionary theory. *Trends in Ecology & Evolution, 11*(11), 453–457.
Fogarty, L., Strimling, P., & Laland, K. N. (2011). The evolution of teaching. *Evolution, 65*(10), 2760–2770.
Fudenberg, D., & Maskin, E. (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica, 54*(3), 533–554.
Gibbard, A. (1982). Human evolution and the sense of justice. *Midwest Studies in Philosophy, 7*(1), 31–46.
Gintis, H. (2007). The evolution of private property. *Journal of Economic Behavior & Organization, 64*(1), 1–16.
Hart, H. L. A. (2012). *The concept of law*. Oxford University Press.
Henrich, J. (2004). Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior & Organization, 53*(1), 3–35.
Henrich, J., & McElreath, R. (2003). The evolution of cultural evolution. *Evolutionary Anthropology: Issues, News, and Reviews, 12*(3), 123–135.
Hirshleifer, D., & Rasmusen, E. (1989). Cooperation in a repeated Prisoners' Dilemma with ostracism. *Journal of Economic Behavior & Organization, 12*(1), 87–106.
Hume, D. (2007). A treatise of human nature: A critical edition. Norton, D.J., & Norton, M.J. (Eds.). Oxford University Press.
Izquierdo, L. R., Izquierdo, S. S., & Sandholm, W. H. (2019). An introduction to ABED: Agent-based simulation of evolutionary game dynamics. *Games and Economic Behavior, 118*, 434–462.
Kameda, T., & Nakanishi, D. (2003). Does social/cultural learning increase human adaptability? Rogers's question revisited. *Evolution and Human Behavior, 24*(4), 242–260.
Kandori, M., Mailath, G. J., & Rob, R. (1993). Learning, mutation, and long run equilibria in games. *Econometrica, 61*(1), 29–56.
Kelly, D., & Setman, S. (2021). The psychology of normative cognition. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* Springer.
Kitcher, P. (1999). Games social animals play: Commentary on Brian Skyrms's evolution of the social contract. *Philosophy and Phenomenological Research, 59*(1), 221–228.
Kline, M. A. (2017). Teach: An ethogram-based method to observe and record teaching behavior. *Field Methods, 29*(3), 205–220.
Laland, K. N., & Hoppitt, W. (2013). *Social learning: An introduction to mechanisms, methods, and models*. Princeton University Press.
Leimar, O., & Hammerstein, P. (2001). Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society B: Biological Sciences, 268*(1468), 745–753.
Lewis, D. K. (1969). *Convention: A philosophical study*. Harvard University Press.
Maynard Smith, J., & Price, G. R. (1973). The logic of animal conflict. *Nature, 246*(5427), 15–18.
McAdams, R. H. (1997). The origin, development, and regulation of norms. *Michigan Law Review, 96*(2), 338–433.
Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature, 437*(7063), 1291–1298.
O'Connor, C. (2019). *The origins of unfairness: Social categories and cultural evolution*. Oxford University Press.
Okasha, S. (2006). *Evolution and the levels of selection*. Clarendon.
O'Connor, C. (2022). Methods, models, and the evolution of moral psychology. In M. Vargas & J. M. Doris (Eds.), *The Oxford handbook of moral psychology* (pp. 441–464). Oxford University Press.
Pettit, P. (2019). Social norms and the internal point of view: An elaboration of Hart's genealogy of law. *Oxford Journal of Legal Studies, 39*(2), 229–258.
Pettit, P. (2023). *The state*. Princeton University Press.
Richerson, P., Baldini, R., Bell, A. V., Demps, K., Frost, K., Hillis, V., Mathew, S., et al. (2016). Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence. *Behavioral and Brain Sciences, 39*, e30.

Richerson, P., & Boyd, R. (2006). *Not by genes alone: How culture transformed human evolution*. University of Chicago Press.
Rogers, A. R. (1988). Does biology constrain culture? *American Anthropologist, 90*(4), 819–831.
Schlag, K. H. (1998). Why imitate, and if so, how? A boundedly rational approach to multi-armed bandits. *Journal of Economic Theory, 78*(1), 130–156.
Shoemaker, D., & Vargas, M. (2021). Moral torch fishing: A signaling theory of blame. *Noûs, 55*(3), 581–602.
Skyrms, B. (1996). *Evolution of the social contract*. Cambridge University Press.
Skyrms, B. (2014). *Evolution of the social contract* (2nd ed.). Cambridge University Press.
Skyrms, B., & Zollman, K. J. S. (2010). Evolutionary considerations in the framing of social norms. *Politics, Philosophy & Economics, 9*(3), 265–273.
Sober, E. (1983). Equilibrium explanation. *Philosophical Studies, 43*(2), 201–210.
Sober, E., & Wilson, D. S. (1998). *Unto others: The evolution and psychology of unselfish behavior*. Harvard University Press.
Sterelny, K. (2012). *The evolved apprentice: How evolution made humans unique*. The MIT Press.
Sugden, R. (1998). Normative expectations: The simultaneous evolution of institutions and norms. In A. Ben-Ner & L. Putterman (Eds.), *Economics, values, and organization* (pp. 73–100). Cambridge University Press.
Sugden, R. (2004). *The economics of rights, co-operation, and welfare*. Palgrave Macmillan.
Taylor, P. D., & Jonker, L. B. (1978). Evolutionary stable strategies and game dynamics. *Mathematical Biosciences, 40*(1), 145–156.
Thomas, B. (1984). Evolutionary stability: States and strategies. *Theoretical Population Biology, 26*(1), 49–67.
Thornton, A., & Raihani, N. J. (2008). The evolution of teaching. *Animal Behaviour, 75*(6), 1823–1836.
Tomasello, M., Kruger, A. C., & Ratner, H. H. (1993). Cultural learning. *Behavioral and Brain Sciences, 16*(3), 495–511.
Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology, 46*(1), 35–57.
Van Lange, P.A.M., Rockenbach, B., & Yamagishi, T. (2014). Reward and punishment in social dilemmas: An introduction. In P. A. M. Van Lange, B. Rockenbach, & T. Yamagishi (Eds.), *Reward and punishment in social dilemmas* (pp. 1–14). Oxford University Press.
Wakano, J. Y., & Aoki, K. (2006). A mixed strategy model for the emergence and intensification of social learning in a periodically changing natural environment. *Theoretical Population Biology, 70*(4), 486–497.
Weibull, J. W. (1995). *Evolutionary game theory*. MIT Press.
Young, H. P. (1993). The evolution of conventions. *Econometrica, 61*(1), 57–84.
Zollman, K. J. S. (2008). Explaining fairness in complex environments. *Politics, Philosophy & Economics, 7*(1), 81–97.