Biol Philos (2018) 33:16
https://doi.org/10.1007/s10539-018-9627-1

Recent trends in evolutionary ethics: greenbeards!

Joseph Heath[1] · **Catherine Rioux**[1]

Received: 15 March 2017 / Accepted: 13 April 2018 / Published online: 19 April 2018

Abstract In recent years, there has been growing awareness among evolutionary ethicists that systems of cooperation based upon "weak" reciprocity mechanisms (such as tit-for-tat) lack scalability, and are therefore inadequate to explain human ultrasociality. This has produced a shift toward models that strengthen the cooperative mechanism, by adding various forms of commitment or punishment. Unfortunately, the most prominent versions of this hypothesis wind up positing a discredited mechanism as the basis of human ultrasociality, viz. a "greenbeard." This paper begins by explaining what a greenbeard is, and why evolutionary theorists are doubtful that such a mechanism could play a significant role in explaining human prosociality. It goes on to analyze several recent philosophical works in evolutionary ethics, in order to show how the suggestion that morality acts as a commitment device tacitly relies upon a greenbeard mechanism to explain human cooperation. It concludes by showing how some early scientific models in the "evolution of cooperation" literature, which introduced punishment as a device to enhance cooperation, also tacitly relied upon a greenbeard mechanism.

Keywords Evolutionary ethics · Reciprocal altruism · Cooperation · Greenbeard · Ultrasociality

Introduction

A common mistake made by philosophers working under the banner of "evolutionary ethics" has been to suppose that the evolutionary science on this question is

✉ Joseph Heath
joseph.heath@utoronto.ca

[1] Department of Philosophy, University of Toronto, Toronto, Canada

settled—that we have an answer to the question, not just how human morality is possible, but even how it evolved. If this were correct, then the only task remaining for the philosopher would be to draw out the normative implications of this body of science. Unfortunately, there is no such scientific consensus. Indeed, the evolution of human ultrasociality—whether it be altruistic or cooperative—is one of the most important *unanswered* questions in the life sciences (Boyd 2006; Gintis and Bowles 2011). This does not mean, of course, that the evolutionary literature is of no importance to moral philosophy, or that it can be disregarded. While it may not provide an answer to the question, what morality is, it nevertheless *constrains the solution space*, when it comes to determining what morality might be. And while we do not yet have an answer to the question of how the human capacity for ultrasociality evolved, several decades of investigation into the question have ruled out a large number of possibilities (Boyd and Richerson 2005). In other words, we know a great deal about how morality did *not* evolve, or what constitutes an implausible answer to this question. Indeed, the fact that the question has remained unanswered for so long is not due to lack of interest, it is rather because most of the solutions proposed over the years have not stood up to scrutiny.

Philosophers who speculate about the evolutionary origins of various human traits are often accused of producing "just so" stories—speculative "adaptive" accounts of particular traits, based on no particular evidence, but that are not easily falsifiable (Kitcher 1985). In the field of evolutionary ethics, however, the problem has been quite the opposite. Rather than producing "just so" stories, most accounts of "the evolution of morality" that have been put forward (e.g. Thompson 1995) rely upon discredited or dubious hypotheses. Typically the mistake lies in a failure to grasp the severity of the free rider problem, which so effectively filters out most forms of altruistic behavior from other sexually reproducing species (e.g. Ruse 1988). This has led many to rely upon mechanisms—such as "tit-for-tat" reciprocity (Axelrod and Hamilton 1981; Axelrod 1984)—that are far too weak to sustain sociality on the *scale* that is observed in humans. The mistake, in other words, lies in taking a mechanism that can explain the extremely modest sociality exhibited by other primates and positing it as an explanation for the vastly more extensive forms of cooperation found in human societies (e.g. de Waal 2006).

One can see this problem most clearly in the philosophical work done under the influence of "first generation" sociobiology, particularly Robert Trivers's (1971) analysis of "reciprocal altruism." (e.g. Ruse 1986). In recent years, there has been growing awareness of the fact that these "weak" systems of reciprocity are not scalable, so that while they may explain cooperation in dyadic or small-group interactions, they cannot explain large-scale cooperation among genetically unrelated individuals (Boyd and Richerson 1988). Thus there has been a trend in recent evolutionary ethics to rely upon models that strengthen the cooperative mechanism, by adding various forms of commitment or punishment. Unfortunately, the most prominent versions of this hypothesis wind up positing yet another discredited mechanism as the basis of human ultrasociality. Philosophers such as Joyce (2006, 2014, 2017) and James (2009, 2011) have both wound up positing what is known as a "greenbeard" to explain human prosociality. This is presumably inadvertent, and indeed, to describe their views in this way is implicitly critical, since the term "greenbeard" is

used primarily as a way of warning against a particular sort of fallacy in evolutionary reasoning. Thus the structure of this paper is simple. We will begin by explaining what a greenbeard is, and why scientists are doubtful that such a mechanism could play a significant role in explaining human prosociality. We then go on to analyze several recent philosophical works in evolutionary ethics, in order to show how the suggestion that morality acts as a commitment device tacitly relies upon a greenbeard mechanism to explain human cooperativeness. By way of consolation, we will also explain how some scientific models in the "evolution of cooperation" literature, which introduced punishment as a device to enhance cooperation, also tacitly relied upon a greenbeard mechanism. Our discussion here follows Gardner and West (2010, 33), who were the first to point out how much of the recent "evolution of cooperation" literature involves confusion over greenbeards.

The narrow objective of the paper is simply to reinforce the idea that "the evolution of cooperation" is a hard problem, and so to the extent that our understanding of morality depends upon its solution, we remain in the dark on certain key issues. There is, however, a broader point, involving philosophical considerations that are often in the background of this debate. Many contemporary religious thinkers defend their convictions using one or another form of ethico-theological argument—claiming, roughly, that the scientific worldview is unable to vindicate the claims of morality, and that this generates a *reductio* of that position (see Lilla 2007, 113–150). Faced with this argument, there is a temptation on the part of evolutionary theorists to respond polemically, by asserting that there is some straightforward scientific account of morality available (e.g. Howson 2011, 137–168; Harris 2010). This temptation should be resisted, simply because the basic claim is untrue. When we affirm our commitment to evolutionary science as an approach to understanding human morality, we are not merely pointing to a body of empirical work that provides an answer to the central philosophical questions, we are committing ourselves rather to a *method of inquiry*, and to a body of empirical work that serves to discipline philosophical speculation in this domain. So far, evolutionary science has not succeeded in providing us with many solutions, when it comes to understanding human morality; its central achievement lies rather in having shown what the important problems are.

What is a greenbeard?

Let us begin with a hyperbolic statement of the problem. If we define *biological altruism* in the standard way, as a property or behavior (collectively, a "trait") that reduces the direct fitness of the individual who possesses it, yet increases the fitness of some other individual (Kerr et al. 2004; Fisher 1930), then it seems almost axiomatic that such a trait cannot evolve by natural selection. It is not just unfit, but doubly so. Consider the first individual born with a desire to risk his own life, in order to save others, or to give up some of her own food, so that others will have enough. Not only does this reduce the fitness of the individual who possesses the trait, it increases the fitness of those who do not! It is difficult to imagine a more obviously lethal mutation. It is, in fact, slightly misleading to say that natural selection has a

tendency to eliminate altruism. As far as the baseline tendency is concerned, there is nothing to eliminate, because altruism cannot arise in the first place—even if it shows up once, by random mutation, it cannot spread in a population. This is why, as Michael Tomasello notes, "altruism is often defined by evolutionary biologists—humorously but pointedly—as 'that which cannot evolve'" (2016, 11). Instead, natural selection will tend to promote *selfishness*, which is to say, traits that increase the direct fitness of the individual that possesses them, while possibly decreasing the fitness of others.

Biological altruism is obviously not the same as altruism in the moral sense, where the notion of interest must be defined, not in terms of fitness, but in terms of the individual's goals. Many people pursue goals that do not, in fact, improve their reproductive fitness, and so the biological and the moral conceptions of altruism are not equivalent, either intensionally or extensionally. Nevertheless, they are far from disjoint. Indeed, if one were to draw a Venn diagram, the area of overlap would be quite large. Many canonical examples of moral action involve spontaneous "helping" behavior, such as jumping into a pool to save a drowning infant, or throwing a switch to divert a runaway trolley, which is altruistic in the biological as well as the moral sense. To the extent that morality prohibits interpersonal aggression and violence, including assault, murder and rape, as well as theft, it is also altruistic in the biological sense. Obligations of generalized benevolence are biologically altruistic as well, insofar as they involve some cost—no matter how small—to the actor. Finally, even if these examples do not exhaust the moral domain, there can be no doubt that whatever traits underlie our capacity to act morally, these traits dramatically increase the probability that an individual will act in a way that is altruistic in a biological sense, and so there is a puzzle about how such traits could have evolved—how they could have "slipped through the net" of natural selection.

There are, of course, certain well-known answers to this question—exceptions to the general rule that natural selection precludes altruism. These are kin selection, reciprocal altruism, and group selection (Gintis and Bowles 2011). The first mechanism relies upon the fact that genes are indifferent between benefits that accrue to themselves and benefits that accrue to copies of themselves that can be found in the environment. Since close relatives are likely to share genes, altruistic traits whose benefits redound primarily to kin may be adaptive, so long as the sum of benefits is sufficiently large that, even when discounted by the coefficient of relatedness of the beneficiaries, it is larger than the cost to the bearer (Hamilton 1964; Gardner et al. 2011). The altruistic trait, in this case, does well because, despite reducing the fitness of the individual who possesses it, the primary beneficiaries are likely to possess the same trait, and so when the benefit to cost ratio is sufficiently high the trait will come out ahead. The second mechanism, reciprocal altruism, occurs when the altruistic trait benefits another individual, at some cost to the bearer, but also has the effect of evoking a response from the other individual that benefits the bearer. As a result, the trait is not really reducing its own bearer's fitness, but is increasing it indirectly. The important difference between this mechanism and kin selection is that reciprocal altruism is able to explain interspecies symbioses—because when the beneficiary is a member of a different species, the copy of the gene found in one individual could not

be benefiting another copy of itself, found in the beneficiary. Thus there must be some mechanism in place through which each one, by providing benefits to the other, is ultimately benefiting itself.

Finally, there is the mechanism of group selection. This has been the subject of much misunderstanding over the years, and so it is important to distinguish modern multilevel selection theory (Wilson 1975, 1977; Colwell 1981; Wilson and Colwell 1981) from earlier, "naive" group selection theories (Wynne-Edwards 1962). The latter were based upon the erroneous assumption that pointing to some collective benefit of a trait was sufficient to provide an adaptive explanation for it. One can still find instances of this in the literature on evolutionary ethics, among authors who appeal to diffuse group benefits, such as "increasing social cohesion" or "reducing lethal violence," as providing an evolutionary explanation for morality (e.g. Ruse 1986). As such, the reasoning is fallacious, since it simply fails to discharge the central burden of proof of any evolutionary explanation, which is to explain how a particular trait manages to reproduce itself more successfully than its rivals in a population. Modern group selection theory, by contrast, is based upon the observation that if a population is segmented into groups, and if the level of genetic variation *between* groups is greater than that *within* groups, then an altruistic trait that increases the fitness of all members of a group can increase its representation in the overall population, even if it is being selected against within each individual group. This is a completely general phenomenon, as witnessed by the fact that kin selection can be represented as a type of multi-level selection (with the family constituting the group) (Marshall 2011; Gardner et al. 2011). There are, however, cases in which the mere partitioning of the population into endogenously reproducing groups (or *demes*) is capable of generating the effect. It is in these cases that is it appropriate to speak of group selection as a *sui generis* mechanism (Woodcock and Heath 2002).

The problem with all three of these mechanisms is that they are unable to account for altruism on the *scale* that one encounters in human societies. The division of labor among humans, for instance, encompasses billions of individuals spread across the entire globe. By contrast, the altruism produced through kin selection, in a sexually reproducing species, is both *partial* and *limited*, as a consequence of the coefficient of relatedness between individuals being no higher than 0.5. It is *partial* because a benefit provided to another never "counts" for more than half of the cost that it imposes upon the actor, and it is *limited* because not just anyone is eligible to become a beneficiary, there is *necessarily* a structure in place that limits the scope of the benefits (even if it is an entirely environmental structure, such as spatial proximity). With reciprocal altruism the problem is that the cooperative structures that can be sustained through mere reciprocity are not scalable. As the number of individuals involved in the interaction increases, the chances that *someone* will defect increases as well. Because the punishment mechanism—withdrawal of cooperation—is not targeted, it does not isolate the individual who defected, but imposes costs on the innocent and guilty alike. This in turn will provoke retaliation from the innocent (if they have not done so already), further intensifying the unravelling of the cooperative system. Thus one defection will set off a cascade of punitive defections, leading the system of cooperation to unravel almost immediately. These observations raise significant doubts about whether reciprocity is the right sort of mechanism

to explain cooperation in the large-scale, anonymous interactions that characterize modern human societies (Boyd and Richerson 1988).

Finally, with group selection, there are serious doubts about whether the empirical conditions required for the effect to be realized can be found among humans. On the one hand, it is the case that human populations in the environment of evolutionary adaptation (EEA) were almost certainly divided up into small groups, and these groups competed with one another fairly intensively. This clearly meets one of the conditions that must be satisfied in order for group selection to be a force. The problem, as Richerson and Boyd have argued (2005), is that there is no evidence of endogenous reproduction and periodic recombination—on the contrary, evidence from contemporary hunter-gatherers suggests significant gene flow between population groups (e.g. exogenous marriage practices, abduction of women in raids, etc.), despite the prevalence of antagonistic social relations. Thus there is no reason to believe that the level of genetic variance between human groups would have been much higher than that within groups (Hill et al. 2011).

Finally, it is worth noting that any explanation for the appearance of ultrasociality among humans must *also* explain its absence in most every other animal species. As a result, the adaptive benefits that are appealed to in any explanation cannot be too obvious—they cannot be the evolutionary equivalent of a $20 bill lying on the ground—otherwise many other species would be taking advantage of them as well. Suppose, for example, that one were to propose the following as an adaptive account of the development of morality: "Morality promotes group cohesion, which in turn generates significant benefits, such as improved defense against predators, more successful hunting, and a reduction in lethal violence within the group." This may be true, but it does not provide any sort of explanation for the emergence of morality. There is the obvious point, that the explanation appeals to diffuse collective benefits, and is therefore an instance of naive group selectionism. But there is also the more subtle point that these collective benefits are ones that would be good for almost *any* species, not just humans. If there were any sort of straightforward connection between these collective benefits and the development of cooperative traits, then one would expect to find high levels of cooperation throughout the animal kingdom. And yet one finds traits such as an advanced division of labor only in very select species, such as the social insects, naked mole rats, and humans (Wilson 2012).

As a result, many theorists who set out to explain human ultrasociality wind up inadvertently undermining their own explanations, by focusing on the continuities in social behavior between humans and other animals, particularly primates. de Waal (2006), for instance, has spent a great deal of time seeking to explain "how morality evolved," and yet he focuses almost entirely on the existence of pro-social behavioral traits among primates, especially chimpanzees. He fails to recognize that, because chimpanzees possess these traits, and yet are incapable of maintaining complex forms of sociality (such as an advanced division of labor), these traits must therefore be *excluded* as an explanation for human ultrasociality. The fact that chimpanzees are capable of sophisticated reciprocal altruism, and live in small, territorial groups, similar to those that humans are thought to have inhabited in the EEA, tends rather to suggest that reciprocal altruism and group selection are insufficient to explain human ultrasociality—otherwise, we would expect to see far more extensive

systems of cooperation among chimpanzees. Thus these discussions of primate sociality, far from explaining how morality evolved, serve only to deepen the mystery. What is needed, in order to solve the "puzzle of human cooperation," is to find something that humans do, that is very similar to what primates do, but that is also in some way qualitatively distinct, such that it is able to sustain a radically different form of sociality (Boyd and Richerson 2005).

It is considerations such as these that have motivated the search for some additional trait, possessed by humans, that might have *potentiated* one of these three mechanism, taking a structure that is capable of producing a modest level of cooperation (e.g. among kin, or in small groups) and somehow expanding it, so that it is able to sustain large-scale cooperation among genetically unrelated individuals.[1] In the case of group selection, for instance, Boyd and Richerson have argued that, because social learning among humans exhibits a conformist bias, within-group variance is likely to be much lower than between-group variance with respect to culturally-transmitted traits, and so group selection will be a more powerful force in the domain of culture than it is in biology (2005). This is what underlies their claim that culture "potentiates group selection" (Boyd and Richerson 2009).

Similarly, with kin selection the search has been on for some mechanism that might allow altruism to emerge among individuals who are not closely related to one another. As Andy Gardner, Stuart West and Geoff Wild have observed, the term "kin selection" is potentially confusing, since the fundamental mechanism at work is one that relies upon "genetical relatedness," not "genealogical relatedness," to sustain altruism (2011, 1024). In other words, what matters in kin selection is not actually kinship, but rather just the probability of possessing the same gene at the particular locus that generates the altruistic trait. Absent some system of reciprocity, an altruistic trait cannot survive and prosper if it directs its benefits primarily toward individuals who do not possess the same trait (i.e. free riders). Thus there must be some correlation mechanism in place, so that the primary beneficiaries of the altruistic trait are others who are likely to possess the same trait. Genealogical relatedness (i.e. being "kin") is one way of achieving this. But if there were some *other* way of identifying, and directing benefits toward, individuals likely to possess the same trait, then that would serve just as well (Hamilton 1964; Gardner et al. 2011, 1026). In order to illustrate the principle, Dawkins (1976) suggested that if a gene led individuals to grow a green beard, and at the same time, disposed them to act altruistically toward other individuals with green beards, then this form of altruism could prosper. The gene would, in effect, have some other way of identifying copies of itself in the population, and of directing the benefits of its altruistic conduct toward these copies.

Dawkins presented this as a hypothetical mechanism, one that he thought was unlikely to be realized, but subsequent research has identified several instances of what is now known as "greenbeard altruism" in nature (West and Gardner 2010). An example of this can be found in the social amoeba *Dictyostelium discoideum*, which possesses an allele (the csA gene) that, under conditions of food scarcity, identifies

[1] This motivation is explicit in Joyce (2006), 40–41. See also Randolph Nesse (2001, 5).

copies of itself in the environment, and joins up with individuals who possess that allele to form cooperative fruiting bodies (Queller et al. 2003). Individuals who lack the allele are excluded—the csA gene produces a cell adhesion protein, so those without it are unable to join with others to form a cooperative group.

Studies have also identified an interesting parallel phenomenon, which is "greenbeard spite" (Hamilton 1970; Gardner and West 2004, 2010). A spiteful trait is one that both reduces the fitness of the individual who possesses it and that of some other individual. The spiteful individual, in other words, engages in behavior that harms others, at some cost to itself (this is to be distinguished from mere selfishness, in which an individual may harm others, but in a way that benefits itself). One might think that spite is pointless, and that it would never develop through natural selection, except that it can be produced through something like negative kin selection. If the spiteful behavior is primarily directed toward *non-kin*, it may be adaptive. Although the trait reduces the fitness of those who possess it, it may reduce the fitness of those who do not possess it *even more*. (Some have suggested that territoriality is a spiteful trait—it is a costly form of aggression, but it survives because it is directed predominantly toward non-kin (Verner 1977; Knowlton and Parker 1979)). One can also find instances of greenbeard spite, in which individuals possess some marker that exempts them from harmful treatment by those who possess the same marker. An example of this are bacteriotocin-encoding genes in many bacteria, which produce both a toxin and a factor that neutralizes the effects of the toxin, so that they kill only those who do not possess the gene (Riley and Wertz 2002; Gardner and West 2010).

Despite these interesting examples, the greenbeard mechanism does suffer from an important limitation. In order to function, it requires three components to be present simultaneously: the production of some sort of overt signal (i.e. the green beard), the capacity to recognize the signal, and the disposition to act altruistically toward those who produce the signal (Queller et al. 2003, 105). The obvious vulnerability of greenbeard altruists is to invasion by "falsebeards," which is to say, individuals who have the green beard, but lack *either* the disposition to help other greenbeards or the capacity to recognize them (in the latter case, the altruistic disposition would be latent—present but never activated). Thus if the three traits are subject to recombination, then they will naturally be subject to variation in the population, which will *necessarily* lead to selection in favor of falsebeards (simply because those who have the green beard, but are slightly-less-disposed to act altruistically toward other greenbeards, will do better than those who are slightly-more-disposed to act altruistically) (Dawkins 1976; Gardner and West 2010). The only way to avoid this is if there is a pleiotropic gene, which is to say, a single genetic locus that produces the signal, the recognition mechanism, and the altruistic behavior—such that it is physically impossible to have one without the other two. This is, however, a condition that is far more likely to be satisfied with micro-organisms, which have rather simple genomes, as well as a much closer link between genotype and phenotype (Foster et al. 2004, 694; Gardner and West 2010, 34). Thus it is no accident that the most convincing cases of greenbeards that have been found in nature are among bacteria, amoebae, yeasts and molds. With the social amoeba, for instance, it is the *same* csA gene that produces both the adhesion to others and the cooperative

behavior, just as with bacteriotocins, the same genetic locus is involved in both the production of the toxin and the factor that deactivates it. With human beings, by contrast, it is seems quite unlikely that a single genetic locus could generate a conditional behavioral disposition *as well as* an observable quality or display, simply because all human behavior is generated by a very complex interplay of many genes, not to mention developmental and environmental factors. Furthermore, human morality seems to involve both a range of signals and a suite of different behaviors, making it difficult to see how all of it could be under the control of a single pleiotropic gene. As a result, the greenbeard mechanism is widely considered to be an implausible basis for an account of human ultrasociality.

Wearing our hearts on our sleeve

Despite this formidable set of objections, the lure of the greenbeard has been strong for evolutionary ethicists. In some cases, the introduction of a greenbeard has been inadvertent, in other cases less so. We would like to begin by discussing an explicit case, found in the moral sentiment theory put forward by Robert Frank, most notably in his book *Passions Within Reason* (1988). This has become something of a touchstone in subsequent discussions of greenbeards, and although the term does not figure in the original presentation of the theory, Frank has subsequently acknowledged that his model relies upon a greenbeard—and has sought to defend it against the usual criticisms. His defense is, we will argue, unpersuasive. It is valuable, however, in that it illustrates some of ways in which ethicists have underestimated the seriousness of the problems confronting greenbeard accounts.

Because several of Frank's key arguments are presented using informal reasoning, it can sometimes be difficult to understand the exact claims being made, or to track the logic of the position. As a heuristic guide, therefore, the following can be proposed as a way of identifying greenbeards. Imagine that you are the altruistic trait. The only way you can survive is by ensuring that the beneficiary of your actions is someone who possesses the same altruistic trait. How do you know who to help? If the way that you determine this is by observing the other person perform an altruistic action, then the mechanism is actually a type of *cooperation* ("direct" reciprocity if you are the beneficiary of the altruistic action, "indirect" reciprocity if someone else is). If the way that you determine it is by deciding that the person is related to you, then the mechanism at work is *genealogical kin selection*. If the way that you determine it is by observing that the individual possesses some other trait that is closely associated with the possession of the altruistic trait (e.g. correlated), then the mechanism is a *greenbeard*. The first two mechanisms are considered respectable, and instances can be found throughout nature. It is the last mechanism that is considered somewhat dubious, and probably nonexistent in the human case.

In Frank's analysis, solving the problem of cooperation involves overcoming a "commitment problem." Consider the variant of a prisoner's dilemma shown in Fig. 1. The only subgame perfect equilibrium of this sequential move game is for player 1 to defect, ending the game right away. He could choose C, giving player 2 the option to move. And while in principle player 2 might also choose C, resulting in

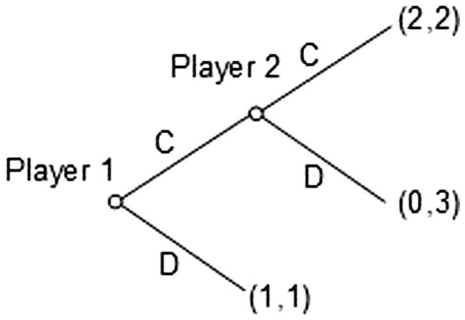

Fig. 1 Commitment problem

the outcome of (2, 2), which is better for both players, in practice player 2 is likely to choose the utility-maximizing option of D, which gives her a payoff of 3, but gives player 1 a payoff of 0. Since player 1 can anticipate player 2′s defection, it is not in his interest to choose C, which is why he will choose to end the game right away by selecting D. This is, of course, frustrating for player 2—if she could only find some way to commit herself, in advance, to playing C, then that might be sufficient to induce player 1 to play C as well, yielding a better outcome for them both. Unfortunately, just announcing an intention to play C in advance of the game would count as "cheap talk"—since it does not affect payoffs, player 1 will ignore it as non-credible.

In order to resolve the problem, player 2 would have to be able to do two things—first, she would need some way of committing herself to playing C, and second, she would need some way of credibly signalling this fact to player 1. Frank argues that the primary function of moral sentiments is that they solve these two problems (Frank 1987, 595). He is somewhat vague about the exact mechanism, but the idea seems to be that certain people experience moral emotions that alter their incentives, in such a way as to make C more attractive to them than D. This is a common theme in moral sentiment theory—because moral emotions are experienced as involuntary they are "hard to fake," and thus provide individuals with a way of making binding commitments. If it all occurred privately, of course, it would be rather unhelpful, because interaction partners would have no way of knowing that the agent was committed (Frank 1987, 594). As a result, Frank claims, the emotional complex that generates the cooperative behavioral disposition also produces an observable signal—such as an involuntary facial display (e.g. blushing in the case of shame). Because the production of the signal is tied to the experience of the emotional state, which in turn is what motivates the cooperative behavior, the signal is *credible*, Frank claims, in a way that a mere statement of intention would not be.

The suspicion that this is a greenbeard gets raised early in the discussion, when Frank asks the reader to "imagine that cooperators are born with a red 'C' on their foreheads, defectors with a red 'D'." (Frank 1988, 59). He goes on to observe that under such circumstances, in which "cooperators and defectors are perfectly distinguishable from each other," the evolution of pro-social behavior would not be subject to any difficulties. The real world, he acknowledges, is not so simple, but it does bear a certain resemblance. Imagine that human cooperators are able to produce a signal that is necessarily conjoined with the activation

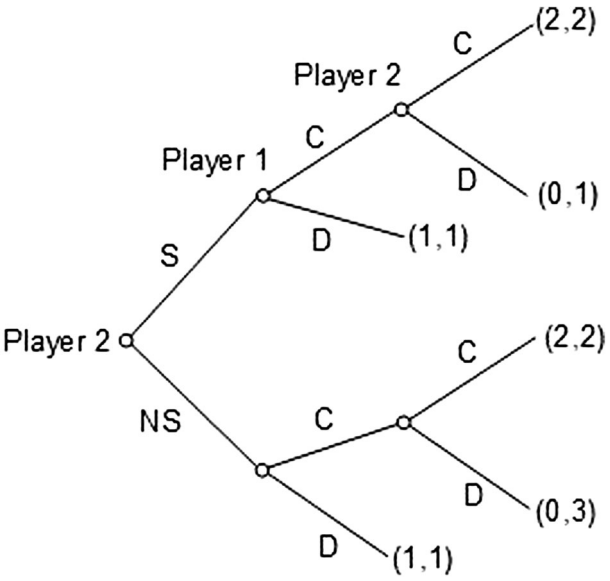

Fig. 2 Commitment problem with signal

of certain moral emotions, which involuntarily dispose the individual to act cooperatively. Then the "commitment problem" would be transformed into something like the game shown in Fig. 2. By producing the signal (S), player 1 essentially modifies his own incentives, so that C is now preferred to D. This allows player 1 to choose C in full confidence that player 2 will do the same. (This is, of course, a simplification, because it assumes that the signal cannot be faked.)

When presented in this way, the mechanism is obviously a greenbeard. There are two distinct traits: "producing the signal" and "acting cooperatively." These are supposed to be correlated with one another, so that if one acts cooperatively toward someone who produces the signal, one is likely to be acting cooperatively toward someone who is similarly disposed toward acting cooperatively. The moral emotion (e.g. the propensity to experience guilt if one does not honor one's promises) is posited as a proximate mechanism explaining the connection between the two traits. Nevertheless, if the two traits are distinct and subject to variation, natural selection will inevitably reduce the connection, because those who produce the signal, yet are less inclined to act cooperatively, will enjoy an advantage over those who are more inclined to act pro-socially. In other words, the population will be invaded by falsebeards (Robson 1990, 387). The suggestion that the "moral emotion" somehow makes this less likely, because it is "hard to fake," is question-begging. If it is hard to fake, then it must have evolved to be that way, and therefore it must be adaptive—but as we have seen, the fact that it precludes free-riding shows that it is *not* adaptive, since those who are able to fake it do better than those who are unable to. It is tempting to think that the moral emotion is adaptive *because it facilitates cooperation,* but that would be to commit the fallacy of naive group selectionism.

Frank is aware of the basic problem, acknowledging that a "mutant strain of defectors" might arise, who produce the signal but not the cooperative behavior. He has two responses to this difficulty. First, he suggests that falsebeards might not arise in the first place, because the signal might be so complicated that defectors would not be able to copy it. "The basic problem for the emergence of strategic signals is that natural selection cannot be forward-looking. It cannot recognize, for example, that a series of mutations might eventually produce an individual able to signal its capacity to solve one-shot social dilemmas, and then favor the first costly step to that end, even though it yields no immediate benefit" (Frank 2005, 90–91). This response, however, misconstrues the problem. Frank appears to be imagining that a falsebeard must arise from a population of individuals who possess neither the capacity to signal nor the disposition to act pro-socially—and thus evolution must reconstruct the signal. Yet the more obvious place for the falsebeard to arise would be from the population of greenbeards, simply through loss of the disposition to act pro-socially or of the capacity to recognize the signal (while retaining the capacity to produce the signal). Frank's argument would be valid if the case involved something like Batesian mimicry, where one species copies another. (The example Frank gives, which suggests that he is thinking along these lines, is that of the Viceroy butterfly imitating the Monarch (Frank 1987, 595).) Here the signal would have to be reproduced from scratch. But within a species, such copying is not required, falsebeards can be created simply through elimination of the cooperative disposition, or the capacity for signal recognition, from a greenbeard.

The same error can be found elsewhere in the literature. Michael Owren and Jo-Anne Backorowski, for instance, have advanced a similar theory of emotional displays, this time focusing on positive rather than negative affect. They suggest that "applying the basic tenets of selfish-gene selection to the evolution of communication in early hominids points to the emergence of honest signaling of positive affect as a valuable mechanism for forming and maintaining cooperative relationships. The same principles dictate that such signals must have emerged in an inherently safeguarded form that provided protection against exploitation by dishonest signalers" (2006, 176–177). Short of pleiotropy, however, there is no such thing as an "inherently safeguarded form" for a signal. Owren and Backorowski are again assuming that falsebeards must find some way of reproducing the signal from scratch: "If smiling and laughter originally arose in a form that ensured authenticity, then these facsimile versions must have arisen at some later time and have different neural foundation" (2006, 158). There is, however, no need for falsebeards to produce a newer, different version of smiling, they need only acquire the capacity to inhibit the cooperative behavior while retaining the original smile.

Frank's second line of defense involves the suggestion that a sort of contest might have arisen, in which defectors begin to copy the signal, but the cooperators respond by modifying the signal, in order to stay one step ahead of the defectors: "Defectors… have no monopoly on the power to adapt. If random mutations alter the cooperators' distinguishing characteristic, the defectors will be faced with a moving target" (Frank 1988, 61). Frank does not develop this suggestion much further, but the idea of a shifting signal is one that has been picked up by other theorists and explored under the rather droll title of "beard chromodynamics." (Jansen and van

Baalen 2006; Traulson and Nowak 2007) The idea, roughly, is that altruists might have either a set of signals at their disposal, or else the ability to invent new signals, and so respond to an invasion of falsebeards by switching to a different beard color (like a club that adopts a new "secret handshake" every week). These models, however, are not particularly robust, and so far represent only a theoretical possibility—there are no clear cases of such a system occurring in nature.

The general problem is that, as Frank himself observes, natural selection is not "forward-looking," and so altruists must rely upon "random mutation" alone to produce, or switch to, a new signal. This is equivalent to saying that greenbeard altruism must "re-evolve" each time a signal is compromised by falsebeards. Such altruism, however, is subject to a startup problem. Because individuals only act altruistically toward those who produce the right signal, there must be at least two of them, benefiting each other, in order for there to be any advantage to producing the new signal. Thus the same random mutation must occur to more than one individual. Furthermore, they must not be related to one another, otherwise the mechanism at work will just be genealogical kin selection. As a result, in order for a group to switch from greenbeard altruism to, say, redbeard altruism, at least two individuals must spontaneously switch over, acquiring a red beard, a capacity to recognize red beards, and a disposition to act altruistically toward those with red beards. Falsebeards, by contrast, have no comparable startup problem. Once the system of altruism is in place, the first individual born with a red beard, but no capacity to recognize, or no disposition to cooperate with, other red beards, immediately does better than the "genuine" redbeards.

Because of this asymmetry, chromodynamic models with genuinely mixed populations and random mutation tend to be dominated by long periods of selfishness, punctuated by relatively short outbursts of altruism, which are eventually discovered, exploited, and destroyed by falsebeards (Traulson and Nowak 2007). The window of opportunity for altruism, in these models, is simply the amount of time that passes between the mutant altruists finding each other and their eventual discovery and destruction by falsebeards. As a result, chromodynamic models that exhibit relatively stable levels of cooperation over time do so by introducing additional factors, which independently increase the level of correlation between altruists. Vincent Jansen and Minus van Baalen, for instance, use a spatial model, in which population viscosity increases correlation (Jansen and van Baalen 2006; Traulson and Nowak 2007, 4). As Gardner and West point out, this essentially transforms it into a genealogical kin selection model (2010, 33). Skyrms (2014, 300) offers a cultural model, in which individuals imitate other, successful individuals. As a result, whenever altruists settle on a new signal, it quickly attracts a large number of imitators. Falsebeards, by contrast, arise through mutation, which gives the altruists a "head start" that stabilizes cooperation in the population.

Skyrms's model illustrates a more general point, which is that the chromodynamic approach seems more plausible in the cultural domain than in the biological (Watson 2012; Skyrms 2010, 120). Because a single individual derives no benefit from switching signals, some kind of coordinated movement away from the older, compromised signal is required, in order for altruism to be sustained at any reasonable level. It is worth observing, however, that the type of proximate mechanism

that would be involved in this seems much different from the moral sentiments that Frank focuses upon. What impressed him about blushing, or the experience of guilt, is precisely that they were not under the direct control of the individual. A cultural model, in which individuals are expected to switch signals as the occasion warrants, would by contrast tend to value flexibility, rather than rigidity. In any case, the work in evolutionary ethics that we are focusing on in this paper all assumes a biological framework, and in this context, a chromodynamic model has essentially nothing to offer above and beyond the traditional greenbeard model.

Recent evolutionary ethics

Despite these evident difficulties, Frank's hypothesis has enjoyed considerable influence in the literature on evolutionary ethics. Perhaps the most important variant of Frank's view has been developed by Richard Joyce, in *The Evolution of Morality* (2006), and elsewhere. The primary difference is that Joyce distances himself from the somewhat simple sentimentalism that informs Frank's account, and seeks instead to provide an account of *moral judgment*. Communication, in Frank's model, consists merely in the expression of a conative state. This is reminiscent of the early sentimentalist claim, that a pronouncement such as "murder is wrong" is analytically equivalent to an expression of disapproval, such as "boo to murder!" (Ayer 1936). The problems with this view are well-known. Joyce's most pressing concern is that Frank's view does not explain what we are doing when we judge other people's actions to be wrong, or censure them for acting contrary to morality (2006, 121). He argues, therefore, that what we have evolved is something like a faculty of moral judgment, which both disposes us to signal approval of cooperative behavior and commits us to acting in conformity with those judgments: "Moral thinking has a distinctive emotional profile: Failure to perform an action that one simply wants to do leads to regret; failure to perform an action that one judges to be morally obligatory leads to guilt—and guilt encourages motivation to restore social equilibrium" (Joyce 2017).

Joyce presents the moral judgment mechanism as the solution to a "commitment problem," although it is somewhat difficult to discern the precise nature of the problem as he sees it. Initially, he speaks of it as though it were simply a matter of individuals suffering from weakness of will when it comes to acting morally. On this view, moral judgment "steps in on occasions when prudence may falter" (Joyce 2006, 113) or else makes "the motivation to cooperate often more reliable" (Joyce 2014, 275). At times Joyce sounds as though this is all that is required in order to provide an account of morality. For instance, he writes that "The hypothesis, then, at its first approximation, is that a judgment like 'That wouldn't be right; it would be reprehensible for me to do that' can play a dynamic role in deliberation, emotion, and desire-formation, prompting and strengthening certain desires and blocking certain considerations from even arising in practical deliberation, thus increasing the likelihood that certain adaptive social behaviors will be performed" (Joyce 2006, 113–114). This does not add anything to our understanding of human sociality, however, because it pertains only to "adaptive" behaviors, whereas the challenge is to

explain how the kind of generalized altruism and extensive cooperation that one finds in human societies could have become adaptive.

Later on, Joyce brings in the signalling function of moral judgment, or the "public nature of moral judgments" (2006, 115), as a way of explaining their benefits. In making a moral judgment publicly, "one signals to others… that one is committed to guiding one's own actions by this moral judgment, that one is not going to pursue X… Others who accept this may consequently alter their actions toward the morally committed individual, choosing him for cooperative ventures—deciding that he is a promising mate, a good trading partner, or simply a valuable member of society" (Joyce 2006, 122). Thus the moral judgment serves as a signal, leading others to act cooperatively toward the individual who produces it, who is, by virtue of having made the judgement, also committed to acting cooperatively. In principle, such an account could explain how large-scale cooperation among unrelated individuals could have evolved. The problem is that it is a greenbeard.[2] Even if it were true that humans were incapable of hypocrisy, evolution would favor individuals capable of making moral judgments, or publicly declaring such judgments, then acting contrary to them (i.e. falsebeards).

In later work, Joyce suggests that the signal may be protected from imitators because it is *costly*.

> Thus moral judgments can function usefully not just as personal commitments, but can be signaled in a way that makes them potential interpersonal commitments. The fact that abiding by moral standards generally involves foregoing short-term profits means that morality can function well as a costly signaling device. When choosing partners for a mutually beneficial cooperative venture, it makes sense to prefer those who can *honestly* signal their willingness to participate. And making signals costly is a way of making them honest, for a sufficiently expensive signal costs the signaler more than the profits that might be reaped through dishonesty. Thus, if one's flourishing or very survival depends on being chosen in cooperative ventures (whether it be as a mate or as a member of a hunting party), it may be adaptive to signal in a costly way one's social virtues (Joyce 2013, 131).

This does nothing to rescue the hypothesis. There are some models in which the prosocial behavior is itself the costly signal, indicating some other quality of the agent (Gintis et al. 2001), but this does not appear to be what Joyce has in mind. If the moral judgment is the signal, it makes no sense to say that "a sufficiently expensive signal costs the signaler more than the profits that might be reaped through dishonesty." Both honest and dishonest individuals produce the same signal, and as a result, incur the same signal costs.[3] Since the benefits of defection are, by hypothesis, larger than the benefits of cooperation, a signal too

[2] This is, I should note, tacitly acknowledged by Joyce, when he describes his account as "supplementing Frank's own account rather than disagreeing with the heart of it" (2006, 122).

[3] Some models "stack the deck" against falsebeards by assuming that their signals are costly, or more costly than those of greenbeards (e.g. Sober 1994, 78). This is unmotivated.

expensive to be produced dishonestly would also be too expensive to be produced honestly. The thought may be, therefore, that producing the signal is not itself costly, but that it *makes* it costly subsequently to defect (as in the Frank model shown in Fig. 2). In this case, the signal would be a greenbeard. Again, in the absence of pleiotropy, there is nothing to prevent the emergence of an individual capable of producing the signal without suffering the costs associated with subsequent defection.

Scott James has defended a view that is quite similar to Joyce's, in his *Introduction to Evolutionary Ethics* and elsewhere. Rather than focusing on moral judgment, however, he argues that it is the capacity to act on moral reasons that solves the commitment problem: "by displaying a commitment to holding oneself accountable, at least in principle, for how one governs oneself in the light of those reasons, one considerably improves one's status as a trustworthy bargaining partner," and thus, presumably increases one's opportunities for cooperation (James 2009, 223). Again, the argument is couched in a great deal of informal reasoning, so it is difficult to discern the precise claim being made. To the extent that the model is more than just the standard appeal to reciprocity, it appears to be a greenbeard. James (2011, 61) also cites with approval another greenbeard proposal, put forward by William Irons, that religiosity, and conspicuous performance of religious ritual, might be a "hard-to-fake signal" of commitment to cooperative action (Irons 2001).

In a later statement of the view, James makes the signaling function of moral reasoning more explicit: "Presenting myself as trustworthy will also serve to attract other cooperative individuals who seek mutual advantage. So if I seek out 'like-minded' individuals and also advertise a disposition to retaliate (e.g. refusing to cooperate after another's defection), I can decrease the chances of being exploited and capitalize on long-term cooperative relationships" (2011, 78). Again, there are elements of James's view that are just a standard repeated-game, conditional cooperation model. To the extent that he extends this, however, by introducing moral reasoning as both a signalling device and a commitment mechanism, then he is relying upon a greenbeard to explain the origins of human morality. If "presenting myself as trustworthy" is all it takes to elicit altruistic behavior from others, then obviously what evolution is going to favor are individuals who are good at presenting themselves as trustworthy, without actually being trustworthy.

Finally, although he is not officially an "evolutionary ethicist," Jesse Prinz has also endorsed a variant of Frank's analysis, in order to explain the evolution of moral sentiments such as guilt. His initial presentation of the argument is straightforwardly fallacious, in that it simply points to the benefits of cooperation as providing an adaptive account of the development of guilt. ("If guilt evolved, that would explain our ability to cooperate, and cooperation has huge advantages. We can hunt, gather and groom better when we cooperate… The payoffs are so great that he guilt-prone mutants do better than their remorseless peers. Guilt gets selected by evolution." (Prinz 2012, 251)) Later on, however, he brings in Frank's analysis, suggesting that these moral emotions help to resolve a commitment problem, along the lines sketched out in Fig. 1. It is by advertising one's propensity to feel guilt that one attracts interaction partners, and persuades them to enter into cooperative relations. Again, this is a greenbeard.

Strong reciprocity

Finally, lest it be suspected that philosophers are uniquely vulnerable to the lure of greenbeards, we would like to discuss one case in which a greenbeard inadvertently slipped into the scientific literature as well. Several instances of this have been drawn attention to by Gardner and West (2010), although they did not provide the details of their analysis. So for those who find their claims less than self-evident, we will elaborate on one of their examples, in this case the model of "strong reciprocity" initially forwarded by (Gintis 2000; Gintis and Bowles 2011).

As mentioned above, "tit-for-tat" style reciprocity models are generally considered insufficient to explain human ultrasociality, because withdrawal of cooperation is too weak a punishment mechanism to motivate compliance in large groups. Once this was recognized, the suggestion was made that, instead of merely withdrawing from cooperation in response to defection, individuals might go after the defector in a spiteful manner, reducing their own payoffs in order to lower those of the defector (Gintis 2000). This suggestion was bolstered by the fact that there are many instances in experimental games of individuals behaving in precisely this fashion (Fehr and Gächter 2002; Fudenberg and Pathak 2009; Gintis and Bowles 2011, 164). The presence of individuals with this disposition—known as "strong reciprocity"—in the population would increase the cost of being a defector. ("A strong reciprocator is predisposed to cooperate with others and punish non-cooperators, even when this behavior cannot be justified in terms of self-interest, extended kinship, or reciprocal altruism" (Gintis 2000, 169).) Now, instead of defectors having to worry about correlation (i.e. getting shut out of the benefits of cooperation, or getting stuck interacting only with other defectors), they would also have to worry about being singled out for spiteful punishment. The thought was that it might only take a small number of these "retaliators" in the population to make a cooperative disposition more attractive than defection. Furthermore, as the number of defectors declined, the costliness of the retaliative disposition would decline as well (since there are fewer instances of defection that call for costly punishment to be inflicted), and so a tipping-point might be reached in which the disposition to cooperate, and to retaliate against those who fail to cooperate, might reach fixation in the population.

This version of the model, however, turned out to be a greenbeard in disguise. There are essentially two "altruistic" dispositions involved, the first is to cooperate, the second is to punish those who do not cooperate. We put the term "altruistic" in quotation marks because, even though this form of punishment is often referred to as "altruistic punishment," it is not really altruistic—while it imposes a cost upon the individual carrying it out, it does not produce any benefits for the individual who is targeted by it. On the contrary, it is costly to that person. Thus the trait is actually spiteful, not altruistic. Of course, if the punishment is directed toward defectors, and thus serves to promote cooperation, that might make it "altruistic," but it is question-begging to build that into the definition of the term. Thus the "strong reciprocity" model, as constructed, has two prima facie malapative traits, an altruistic disposition to act cooperatively, and a spiteful disposition to punish those who defect.

Given this initial setup, it is not difficult to see how the altruistic disposition is sustained. But what sustains the punishment system? In order for it to be sustained, it would have to be the case that spiteful punishment is directed toward all those who do not have the disposition to engage in spiteful punishment. Unfortunately, strong reciprocity does not pick out the disposition to engage in spiteful punishment directly, it instead uses the disposition to cooperate as a proxy for the disposition to punish. In this respect, "cooperating" is like the green beard, which is supposed to indicate the presence of the spiteful disposition, "punishing." Because of this, populations of strong reciprocators are vulnerable to invasion by falsebeards, who in this case are the mere cooperators, or weak reciprocators, disposed to cooperate but having no desire to engage in spiteful punishment. They receive all the advantages of cooperation, as well as all the advantages that flow from the relative paucity of defectors, but without shouldering any of the costs associated with punishing these defectors.

This is not necessarily fatal to the strong reciprocators, but it does significantly weaken the model. As Gintis and Bowles later showed, an equilibrium can be reached between strong reciprocators, cooperators and selfish agents, where each keeps the other in check. The strong reciprocators are effective at controlling the selfish agents, but are vulnerable to invasion by cooperators. The cooperators, however, are vulnerable from the selfish agents. Thus there is a polymorphic or mixed equilibrium that involves a balance between all three. While this is of some theoretical interest, the persistence of a very large group of unconditional defectors raises doubts about the usefulness of the model at explaining human ultrasociality. It is also difficult to imagine plausible empirical circumstances in which a population of defectors could be successfully invaded and a stable polymorphic/mixed equilibrium achieved.

Conclusion

The evolution of human ultrasociality remains one of the most controversial questions in the life sciences. Yet while there is no definitive solution to the puzzle, several decades of active debate in the scientific community have succeeded in dramatically narrowing the solution space, providing us with a much better sense of what we are looking for, and which proposed mechanisms are more or less plausible. To the extent that morality is part of the puzzle of human ultrasociality, this narrowing of the solution space has important consequences for moral philosophy. Even though we do not know how human morality evolved, we are in a position to say a great deal about how it could *not* have evolved. The scientific literature has in fact identified several pitfalls for the unwary, which moral philosophers should be conscious of, if they hope to produce theories that can be integrated into a broader scientific worldview. We are also in a position to specify a set of desiderata, that any plausible account of morality must satisfy. One of the most important desiderata is that the theory should not depend upon a greenbeard mechanism for its account of altruism (or if it does, it should provide a plausible account of how it can discharge the rather extraordinary burden of proof associated with such an hypothesis). Unfortunately,

some of the most prominent recent work done in evolutionary ethics fails to satisfy this desideratum.

References

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

Axelrod R, Hamilton WD (1981) The evolution of cooperation. Science 211:1390–1396

Ayer AJ (1936) Language, truth and logic. Victor Gollancz, London

Boyd R (2006) The puzzle of human sociality. Science 314:1555–1556

Boyd R, Richerson PJ (1988) The evolution of reciprocity in sizable groups. J Theor Biol 132:337–356

Boyd R, Richerson PJ (2005) Solving the puzzle of human cooperation. In: Levinson S, Jaisso P (eds) Evolution and culture. MIT Press, Cambridge, pp 105–132

Boyd R, Richerson P (2009) Culture and the evolution of human cooperation. Philos Trans R Soc B Biol Sci 364:3281–3288

Colwell RK (1981) Group selection is implicated in the evolution of female-biased sex ratios. Nature 290:401–404

Dawkins R (1976) The selfish gene. Oxford University Press, Oxford

de Waal F (2006) Primates and philosophers: how morality evolved. Princeton University Press, Princeton

Fehr E, Gächter S (2002) Altruistic punishment in humans. Nature 415:137–140

Fisher RA (1930) The genetical theory of natural selection. Clarendon Press, Oxford

Foster KR, Shaulsky G, Strassmann JE, Queller DC, Thompson CRL (2004) Pleiotropy as a mechanism to stabilize cooperation. Nature 431:693–696

Frank RH (1987) If homo economicus could choose his own utility function, would he want one with a conscience? Am Econ Rev 77:593–604

Frank RH (1988) Passions within reason. Norton, New York

Frank RH (2005) Altruists with green beards: still kicking? Anal Kritik 27:85–96

Fudenberg D, Pathak PA (2009) Unobserved punishment supports cooperation. J Pub Econ 94:78–86

Gardner A, West SA (2004) Spites and the scale of competition. J Evolut Biol 17:1195–1203

Gardner A, West SA (2010) Greenbeards. Evolution 64:25–38

Gardner A, West SA, Wild G (2011) The genetical theory of kin selection. J Evolut Biol 24:1020–1043

Gintis H (2000) Strong reciprocity and human sociality. J Theor Biol 206:169–179

Gintis H, Bowles S (2011) A cooperative species: human reciprocity and its evolution. Princeton University Press, Princeton

Gintis H, Smith EA, Bowles S (2001) Costly signaling and cooperation. J Theor Biol 213:103–119

Hamilton WD (1964) The genetical evolution of social behavior. I & II. J Theor Biol 7:1–52

Hamilton WD (1970) Selfish and spiteful behaviour in an evolutionary model. Nature 228:1218–1220

Harris S (2010) The moral landscape. Free Press, New York

Hill KR, Walker RS, Božičević M, Eder J, Headland T, Hewlett B, Hurtado M, Marlowe F, Wiessner P, Wood B (2011) Co-residence patterns in hunter-gatherer societies show unique human social structure. Science 6022:1286–1289

Howson C (2011) Objecting to god. Cambridge University Press, Cambridge

Irons W (2001) Religion as a hard-to-fake sign of commitment. In: Nesse RM (ed) Evolution and the capacity for commitment. Russell Sage, New York, pp 209–290

James SM (2009) The caveman's conscience: evolution and moral realism. Australas J Philos 87:215–233

James SM (2011) An introduction to evolutionary ethics. Wiley-Blackwell, Malden

Jansen VAA, van Baalen M (2006) Altruism through beard chromodynamics. Nature 440:663–666

Joyce R (2006) The evolution of morality. MIT Press, Cambridge

Joyce R (2013) Ethics and evolution. In: LaFollette H, Persson I (eds) Blackwell guide to ethical theory. Wiley Blackwell, Oxford, pp 123–147

Joyce R (2014) The origins of moral judgment. Behaviour 151:261–278

Joyce R (2017) Human morality: from an empirical puzzle to a metaethical puzzle. In: Ruse M, Richards R (eds) Cambridge handbook of evolutionary ethics. Cambridge University Press, Cambridge

Kerr B, Feldman MW, Godfrey-Smith P (2004) What is altruism? Trends Ecol Evolut 19:135–140

Kitcher P (1985) Vaulting Ambition. MIT Press, Cambridge

Knowlton N, Parker GA (1979) An evolutionary stable strategy approach to indiscriminate spite. Nature 279:419–421
Lilla M (2007) The stillborn god. Alfred A. Knopf, New York
Marshall JA (2011) Group selection and kin selection: formally equivalent approaches. Trends Ecol Evolut 26:325–332
Nesse R (2001) Natural selection and the capacity for subjective commitment. In: Nesse R (ed) Evolution and the capacity for commitment. Russell Sage Foundation, New York, pp 1–45
Owren MJ, Bachorowski JA (2006) The evolution of emotional expression: a "selfish-gene" account of smiling and laughter in early hominids and humans. In: Mayne TJ, Bonanno GA (eds) Emotions: current issues and future directions. The Guilford Press, New York, pp 152–191
Prinz JJ (2012) Beyond human nature: how culture and experience shape the human mind. Norton, New York
Queller DC, Ponte E, Bozzaro S, Strassman JE (2003) Single-gene greenbeard effects in the social amoeba dictyostelium discoideum. Science 299:105–106
Richerson PJ, Boyd R (2005) Not by genes alone. University of Chicago Press, Chicago
Riley MA, Wertz JE (2002) Bacteriocins: evolution, ecology and application. Ann Rev Microbiol 56:117–137
Robson AJ (1990) Efficiency in evolutionary games: darwin, nash and the secret handshake. J Theor Biol 144:379–396
Ruse M (1986) Evolutionary ethics: a phoenix arisen. Zygon 21:95–112
Ruse M (1988) Evolutionary ethics: healthy prospect or last infirmity? Can J Philos 18(sup1):27–73
Skyrms B (2010) Signals. Oxford University Press, Oxford
Skyrms B (2014) Social dynamics. Oxford University Press, Oxford
Sober E (1994) The primacy of truth telling and the evolution of lying. In: Sober E (ed) From a biologicalpoint of view: essays in evolutionary philosophy. Cambridge University Press, Cambridge, pp 71–92
Thompson P (ed) (1995) Issues in evolutionary ethics. SUNY Press, Albany
Tomasello M (2016) A natural history of human morality. Harvard University Press, Cambridge
Traulson A, Nowak MA (2007) Chromodynamics of cooperation in finite populations. PLoS ONE 2(3):e270
Trivers R (1971) The evolution of reciprocal altruism. Q Rev Biol 46:35–57
Verner J (1977) On the adaptive significance of territoriality. Am Nat 111:769–775
Watson E (2012) The evolution of tag-based cooperation in humans: the case for accent. Curr Anthropol 53:588–616
West S, Gardner A (2010) Altruism, spite and greenbeards. Science 327:1341–1344
Wilson DS (1975) A theory of group selection. Proc Ntl Acad Sci 72:143–146
Wilson DS (1977) Structured demes and the evolution of group advantageous traits. Am Nat 111:157–185
Wilson EO (2012) The social conquest of earth. W. W. Norton, New York
Wilson DS, Colwell RK (1981) The evolution of sex ratio in structured demes. Evolution 35:882–897
Woodcock S, Heath J (2002) The robustness of altruism as an evolutionary strategy. Biol Philos 17:567–590
Wynne-Edwards VC (1962) Animal dispersion in relation to social behaviour. Oliver and Boyd, Edinburgh