Biol Philos (2015) 30:405–421
DOI 10.1007/s10539-015-9486-y

Evolution and the classification of social behavior

Patrick Forber[1] · **Rory Smead**[2]

Received: 13 August 2014 / Accepted: 12 March 2015 / Published online: 2 April 2015

Abstract Recent studies in the evolution of cooperation have shifted focus from altruistic to mutualistic cooperation. This change in focus is purported to reveal new explanations for the evolution of prosocial behavior. We argue that the common classification scheme for social behavior used to distinguish between altruistic and mutualistic cooperation is flawed because it fails to take into account dynamically relevant game-theoretic features. This leads some arguments about the evolution of cooperation to conflate dynamical scenarios that differ regarding the basic conditions on the emergence and maintenance of cooperation. We use the tools of evolutionary game theory to increase the resolution of the classification scheme and analyze what evolutionary inferences classifying social behavior can license.

Keywords Evolution · Cooperation · Hamilton's rule · Social behavior · Evolutionary game theory

From slime molds to humans, navigating social life is a crucial challenge—reproductive success turns on the ability to form coalitions, interact with the right individuals, track reputation, and negotiate divisions of labor. The interplay of costs and benefits from these interactions influences the evolutionary fate of social traits.

Patrick Forber and Rory Smead have contributed equally to this work.

✉ Patrick Forber
patrick.forber@tufts.edu

Rory Smead
r.smead@neu.edu

[1] Department of Philosophy, Tufts University, Miner Hall, 14 Upper Campus Rd, Medford, MA 02155, USA

[2] Department of Philosophy and Religion, Northeastern University, Holmes Hall, 360 Huntington Ave, Boston, MA 02115, USA

Table 1 The standard classification scheme for social behaviors based on fitness effects on the actor and recipient

Effect		Recipient	
		Positive	Negative
Actor	Positive	Mutual benefit	Selfishness
	Negative	Altruism	Spite

The aim of this paper is to critically evaluate a common scheme of classifying social behavior based on costs and benefits from a dynamical, game-theoretic perspective, and propose a more sophisticated way to classify social behavior that is sensitive to the strategic options as well as the costs and benefits of behavioral strategies.

The standard and the shift

Classifying social behaviors in terms of their effects on fitness helps structure investigations into the evolutionary origins of such behavior. The seminal work of Hamilton (1964a, b) provides the standard scheme for classification given in Table 1 (West et al. 2007).[1]

Hamilton aimed to distinguish altruistic from selfish behaviors in order to explore when the former might evolve. The implication here is that the standard scheme helps structure important evolutionary generalizations about social behavior. Each class of behavior highlights important general features of the evolutionary processes that tend to produce that behavior. Consider the typical gloss given to each of the four classes of social behavior. Selfish behaviors pose no special evolutionary puzzle since such self-interested behavior generates a benefit for the actor and imposes a cost on the recipient. All else being equal, such behavior should evolve by individual level natural selection. In contrast, altruistic behaviors, traits that involve an actor paying a cost to confer a benefit on a recipient, are particularly puzzling. Individual level selection should never favor behavior that imposes a cost to confer a benefit on a competitor; altruism is vulnerable to subversion by free riders. Evolutionary theory now encompasses a number of models that delineate ways that altruistic behavior can evolve, invoking group selection, population structure, kinship, or conditional behavior. These factors can coordinate interactions such that altruists tend to receive more of the benefit generated by altruistic behavior than free riders (for review see, e.g., Skyrms 1996; Frank 1998; Sober and Wilson 1998; Okasha 2007; Sterelny et al. 2013). Spiteful behaviors, traits where an actor pays a cost to impose a cost on a recipient, are equally puzzling. Similar factors can

[1] Hamilton (1964a, 15) presents the 2 × 2 scheme in terms of inclusive fitness gains and losses; these are now referred to as benefits and costs. In his classification scheme, behaviors where individuals (actors) lose and neighbors (recipients) gain are altrustic. Conversely, behaviors where actors gain and recipients lose are selfish. Interestingly, the other two categories are labeled "selected" (mutual benefit, both gain) and "counter-selected" (spite, both lose); subsequent work revealed a flaw in these labels (Hamilton 1970).

coordinate interactions such that spiteful types tend to inflict costs on non-spiteful types in a way that generates a relative advantage for spite (Hamilton 1970; Price 1970). The last category, mutually beneficial behaviors, traits where both actor and recipient gain a benefit, do not seem to pose any evolutionary puzzle, at least not in the way that altruism and spite do. Individual level selection should favor mutualistic interactions. Indeed, this seems to be the best case for cooperative social behavior.[2]

Notice that the standard scheme is based on the assessment of fitness costs and benefits associated with the actor's behavior. This raises a family of conceptual issues. Of course, there are the familiar problems regarding the measurement of fitness (Taylor et al. 1974; Beatty and Finsen 1989; Beatty 1992; Sober 2001; Ariew and Lewontin 2004); we do not have solutions to present here and will set these problems aside. Rather, we want to flag a different conceptual point: the assessment of costs and benefits (or gains and losses, in Hamilton's terminology) depends on a comparison, often left implicit. The obvious comparison behind Hamilton's standard scheme is that between a status quo baseline and exhibiting the behavior. The costs and benefits of some specific behavior are determined by assessing the fitnesses of actor and recipient had the actor exhibited that behavior versus their fitnesses had the actor not exhibited the behavior at all. With respect to altruism, the implicit strategy space contains two options: the actor behaves altruistically or not. The assessment does not depend on the behavioral strategy of the recipient. In part, the influence of this standard scheme for classification framed the evolutionary problem of cooperation in terms of altruism.

Recent approaches to cooperation, especially human cooperation, have shifted focus from altruism to mutual benefit.[3] While these new approaches vary considerably in their evolutionary hypotheses, they share two unifying features: (1) the discounting of the traditional problems of free-riding and subversion associated with biological altruism; and (2) an emphasis on the problems of coordination and coalition formation.

Sterelny (2012), as part of his sophisticated account of human cognitive evolution, makes this shift, at least for early hominin evolution. In contrast to the Machiavellian social intelligence hypothesis—that an evolutionary arms race between defection and detection drove human cognitive evolution—Sterelny argues for the importance of information sharing and coordinated cooperative foraging: "the cognitive problem of effective coordination is *more demanding* than that of defection" (2012, 10). To coordinate foraging individuals need to pool information. Sharing information to facilitate resource gathering provides significant benefits and minimizes risks of cheating or defection. Thus, the key evolutionary problem becomes one of coordination, not subversion. To put the contrast another way, a coordination

[2] That said, it turns out that even in the so-called easy cases for cooperation certain difficulties can arise in the form of spite Forber and Smead (2014).

[3] Mutually beneficial social behavior is often referred to as "mutualism" or as "by-product mutualism" in the literature. West et al. (2007) argue that the term "mutual benefit" is better than "mutualism," given the common use of the latter term in ecology to describe cross-specific symbioses. Since we do not consider symbioses here, we will use the terms interchangeably, though we will follow this recommendation as best we can.

problem concerns the origination of cooperative behavior in a population. In such scenarios, once the behavior evolves there is no incentive to defect or free-ride and cooperation is stable. A subversion problem concerns the instability of cooperative behavior in a population due to the incentive to defect; invading free-riders can outperform the cooperative natives in such populations. Implicit in Sterelny's argument, that coordination rather than subversion matters, is the commitment that early cooperative behavior was mutually beneficial, not altruistic.

To be clear, Sterelny does not claim that the risks of defection and deception disappear, nor that altruism has no role to play in the evolutionary explanation of human cooperation. Rather, these come into play later in the human lineage: "in comparison to ancient forager lives, cooperation in modern forager life is probably more fragile because it is more optional ...the evolution of explicit and articulated norms is in part a sign of stress on cooperative defaults" (Sterelny 2012, 91). For ancient foragers, at least, cooperation was not classically altruistic.

Tomasello and collaborators make the shift explicitly:

> Our starting point is not cooperation as altruistic helping, but rather cooperation as mutualistic collaboration. Our hypothesis, which we call the Interdependence Hypothesis, is that at some point humans created lifeways in which collaborating with others was necessary for survival and procreation (and cheating was controlled by partner choice) (Tomasello et al. 2012).

In particular, they stress early and increasingly obligate collaborative foraging as setting the stage for the evolution of human cooperation. The emergence of foraging bands increases the interdependence of members by generating mutually beneficial rewards that far outweigh what solo foraging might produce. The interdependence makes free-riding less of a problem, and instead poses cognitively demanding coordination problems that require a kind of shared intentionality to execute. Free-riding, if it ever occurs, can be sanctioned by exclusion from the (mutually beneficial) foraging endeavors.

Some take these arguments a step further, and claim mutualistic cooperative behavior helps explain the evolution of morality (Baumard et al. 2013; Tomasello and Vaish 2013). Since mutualistic cooperation can easily evolve, it readily facilitates varied and complex prosocial behavior. If individuals have the option of choosing their collaborators then so much the better—the incentives to be a cooperative partner are further increased and there may even be competition to interact with the best cooperators. This avid prosocial behavior, so the story goes, leads to the emergence of norms of fairness. These norms get deployed across social situations, whether with strangers or kin. Such fairness norms provide the foundation for proto-morality and the subsequent emergence of fully-fledged moral norms.

It is important to distinguish two features of this family of approaches. First, there is the obvious shift away from costly helping behavior—the standard sort of biological altruism that involves paying a cost to help a partner or one's coalition—as the model of cooperation. In game theoretic terms, the shift is best characterized as trading in the Prisoner's Dilemma, where cooperation is the dominated strategy, for the Stag Hunt, where cooperation is difficult to achieve but stable once it

evolves.[4] This shift can be understood as an alternative interpretation of the central problem of cooperation, focusing on coordination rather than subversion.

The second feature is the emphasis on a particular mechanism for structuring interactions: partner choice. This sort of mechanism can play a role in any strategic interaction and can reinforce cooperative behavior. The ability to choose interaction partners or coalitions can increase the stability and chance of invasion in both Prisoner's Dilemma and Stag Hunt scenarios, provide a way to sanction free-riding should it occur in repeated interactions, and open the door for reputation to play a role in the choice of interaction partners. In short, the introduction of partner choice enriches and complicates the evolution of social behavior by introducing a new strategy to the standard games, that of opting out.

While there is much of value in recent work on mutualistic cooperation, the evolutionary arguments about the origin and stability of mutually beneficial behaviors tend to conflate dynamical scenarios that differ regarding the conditions on how cooperation can emerge and how it can be maintained. In particular, the standard scheme fails to adequately account for game-theoretic and frequency dependent effects that are common in social interactions. Crucially, social behavior that benefits both parties may only do so if the other individuals are also behaving in similar ways. Consider cooperative hunting—it does not pay to hunt game that requires a large hunting party, if all your potential interaction partners are off hunting solo. In such a case, a particular social behavior may be classified as mutual benefit in some interactions, but as altruism (or perhaps even as spite) in others. Coordination is an evolutionary problem for cooperative hunting. This is not so for a behavior that is beneficial to all involved independently of the behavior of others. Such behavior counts as mutualistic, too, but is not subject to similar obstacles in its evolution. The broad category of mutualistic cooperation does not recognize the difference between cases of invariant benefit-to-all and cases where the fitness ordering of strategies is frequency dependent. Due to the importance of frequency dependence, any classification scheme that neglects such effects is unlikely to elucidate evolutionary obstacles or avenues for prosocial behavior.

This same line of reasoning also applies to less cooperative social interactions. Territory contests are cases where aggressive behavior will be classified as selfish in some interactions, and spiteful in others. Furthermore, recent work on the evolutionary dynamics of social interactions reveals both interesting connections and important distinctions between various social behaviors (Lehmann et al. 2006; West and Gardner 2010; Smead and Forber 2013). These observations suggest that classifying social behaviors will be more accurate and relevant if it is done in a way that captures such strategic features.

Classifying social behavior: a test case

To demonstrate the limitations of the standard scheme for classifying social behavior, consider a very simple interaction: the Stag Hunt game (Table 2). In this game, we might imagine two organisms faced with an opportunity to cooperatively

[4] In fact, Tomasello et al. (2012, 674) specifically endorse this game-theoretic characterization of the shift, though they do not pursue their arguments with formal evolutionary game theory.

Table 2 A Stag Hunt game

	Behavior X	Behavior Y
Behavior X	3, 3	0, 2
Behavior Y	2, 0	1, 1

Suppose the actor chooses the row and the recipient chooses the column. Payoffs in each cell are listed for (actor, recipient)

hunt. An organism might behave in a way that fulfills their part of the hunt (behavior *X*), or the organism may neglect their role and seek food on their own (behavior *Y*). The best result for each organism would come from both doing *X*, the worst from doing *X* while the other does *Y*. Doing *Y* while the other does *X* yields a modest reward and doing *Y* while the other does *Y* yields a smaller reward due to potential competition over limited alternative food sources.

How should we classify behaviors *X* and *Y* in this simple two person game? The answer depends on the context in which the behaviors occur. More precisely, whether an actor's behavior counts as mutually beneficial (for example) depends on what the recipient is doing. There is no simple baseline for comparison since the payoff for the actor is determined by the strategy of the recipient (and vice versa). Consider first the actor's behavioral strategy *X*. If the recipient is also exhibiting behavior *X* then the actor's behavior generates a fitness payoff of 3 for the actor and 3 for her recipient. In contrast, the actor's behavior *Y* generates a payoff of 2 for the actor and 0 for her recipient. Opting for *X* rather than *Y* looks mutualistic; both actor and recipient do better than the alternative. However, if the opponent is exhibiting behavior *Y*, then doing *X* generates 0 rather than 1 for the actor, and 2 rather than 1 for her recipient. In this context, the behavior looks altruistic; the actor pays a cost (receiving 0 rather than 1) to confer a benefit on the recipient (she receives 2 rather than 1). The behavioral strategy *Y* presents a symmetric difficulty. If the recipient exhibits *X*, behavior *Y* looks spiteful; the actor pays a cost to inflict a cost on her recipient. Yet if the recipient exhibits *Y*, then behavior *Y* looks selfish. When interactions take the form of the Stag Hunt, possible behaviors can be interpreted in any of the four standard categories depending on the context.[5]

Perhaps more striking, the classification of behaviors *X* and *Y* can change by shifting the payoffs in the same sort of (symmetric) game. Consider a modified Stag Hunt game (Table 3). Cooperation is still stable but now lack of coordination is more severely penalized. The classification of behaviors changes in the modified game. When the recipient exhibits *Y* behavior *X* now looks spiteful (versus altruistic in the original game) and behavior *Y* now looks mutualistic (versus selfish in the original game).[6]

[5] Perhaps more problematic, there are scenarios when classification fails completely. If one's opponent plays a strategy corresponding to the mixed Nash equilibrium (for the above game the mixed strategy exhibits *X* and *Y* with equal probability) then there are no expected costs or benefits associated with *X* and *Y* for the actor.

[6] Thanks to Marty Barrett for making this point clear to us. The details underlying the change in classification are as follows. The payoff comparison for actor's behavior *X* (when recipient does *Y*) in the modified game is 0 versus 2 for the actor and 1 versus 2 for her recipient. Opting for *X* over *Y* inflicts costs on both and so looks spiteful. The payoff comparison for actor's behavior *Y* (when recipient does *Y*) changes in a similar way (2 versus 0 for the actor and 2 versus 1 for her recipient). Opting for *Y* over

Table 3 A modified Stag Hunt game

	Behavior X	Behavior Y
Behavior X	3, 3	0, 1
Behavior Y	1, 0	2, 2

Suppose the actor chooses the row and the recipient chooses the column

This attempt at classification of behavior in the Stag Hung game tells one very little about what to expect regarding the evolutionary dynamics in this game. In traditional population models of evolutionary game theory—infinite, randomly mixing populations playing one-shot games—*X* will evolve if a sufficient number individuals are playing *X*; otherwise *Y* will evolve. It does not help to know whether, given the context of a particular interaction, a behavior is labeled as "altruistic", "mutualistic", "spiteful", or "selfish." What maters is the nature of the *interaction* and the composition of the population. The relevant units of classification are not the social *behaviors*, but the types of social *interactions*.[7] In the case of the Stag Hunt the units are the four possible pairs of behavioral strategies. This shift, from behaviors to interactions, points towards a better classification scheme.

A game-theoretic scheme

The idea of classifying social interactions rather than behaviors can be generalized using standard tools from evolutionary game theory. Here we will illustrate how this can be done for any 2×2 symmetric interaction (Table 4).

Suppose $a > d$. Then we can label for four general classes of strategic interactions:

- Social Delight (e.g., Prisoner's Delight): $a > b$, $c > d$
- Social Dilemma (e.g., Prisoner's Dilemma): $b > a$, $d > c$
- Conflict (e.g., Hawk–Dove): $b > a$, $c > d$
- Coordination (e.g., Stag Hunt): $a > b$, $d > c$

The evolutionary dynamics are primarily governed by the differences between the payoffs, $a - b$ and $c - d$, and the frequency of behaviors in the population. Under the assumption of a single, infinite, and randomly mixing population, these types of

Footnote 6 continued
X benefits both and so looks mutualistic. In asymmetric games different shifts in payoffs are possible, illustrating further flexibility in classifying behaviors.

[7] This way of describing the problem of classification has not, to our knowledge, been made explicit in the literature, though there are some approaches that do focus on the interactions rather than individual strategies. For example, Bomze (1983) provides a "complete classification of the two-dimensional phase flows" for the replicator dynamics where different cases are individuated based differences in the corresponding payoff matrix. The payoff matrices provide the necessary information to classify types of interactions.

Table 4 A 2 × 2 game

	Behavior X	Behavior Y
Behavior X	a, a	c, b
Behavior Y	b, c	d, d

games produce different evolutionary scenarios. For Social Delight games ($a > b$ and $c > d$) behavior X is strictly dominant and will evolve from any initial conditions where both behaviors are present in the population. For Social Dilemma games ($b > a$ and $d > c$) behavior Y is strictly dominant. For Coordination games ($a > b$ and $d > c$) the evolutionary system is bistable and one of the two behaviors will evolve depending on the initial conditions and the exact payoffs in the game. For Conflict games ($b > a$ and $c > d$) the population will converge to a polymorphic mixture of X and Y. This allows us to create a space of possible 2 × 2 interactions where the category of social interaction is representative of the underlying evolutionary dynamics (Figure 1).[8]

The game-theoretic framework encompasses the standard scheme as a special case: it is simply a 2 × 1 game where the actor (row player) can affect the outcome and the recipient (column player) cannot (Table 5). When $a > b$ and $c > d$ behavior X is mutualistic and Y is spiteful. When $b > a$ and $d > c$ behavior X is selfish and Y is altruistic. Inverting each pair of inequalities inverts the classification for X and Y respectively. The standard scheme of classification is simply a rather narrow special case for the more general game theoretic approach (given by the dashed diagonal in Fig. 1).

Of course, if we were committed to something like the standard scheme, it would be possible to simply fix an interaction context and proceed as though it is a 2 × 1 interaction.[9] While this move could recover the standard categories for social behavior, such classifications would have little to no connection with the evolution of those behaviors. In addition, there is no reason to think that classifications of social behavior will remain constant as both the behavior of interaction partners and the composition of the population in which the interaction takes place may change.

On the other hand, if we want to have categories that track the all relevant features of the evolutionary dynamics underlying social behavior, we would need to look beyond 2 × 2 symmetric games. The approach above could be generalized to asymmetric interactions or games with more than two strategies. However, doing so would greatly increase the number of potential interaction categories. Bomze (1983)

[8] Hauert et al. (2006) show how this sort of framework can be extended to N-players and can incorporate synergy and discounting amongst the contributions of multiple players. Bomze (1983) investigates a complex dynamic classification for 3 × 3 symmetric game. For our purposes, the 2-player 2 × 2 framework found in Weibull (1995) suffices to reveal the limitations in the standard scheme, and so we pursue our argument without these additional formal details.

[9] Fixing the interaction context could mean examining a particular interaction with an opponent that is behaving in a particular way. It could be specified in other ways as well. For instance, one might take the distribution of behaviors in a given population and imagining an interaction with an idealized "average" interaction partner.

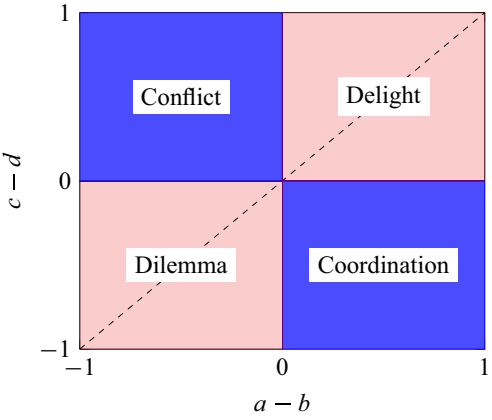

Fig. 1 The space of possible 2 × 2 symmetric games (adapted from Weibull 1995, 28–30). Assume $a > d$. The *dashed diagonal* represents the special case on which the standard classification scheme tracks evolutionary dynamics

Table 5 A 2 × 1 game

	Behavior Z
Behavior X	*a*, *c*
Behavior Y	*b*, *d*

presents a classification scheme for 3 × 3 symmetric games that reveals at least 47 qualitatively different evolutionary scenarios. As interactions become more complicated, classification will become significantly more difficult and, most likely, less helpful for drawing conclusions of evolutionary significance. The categories for social behavior would reveal nothing that we could not glean from simply examining the evolutionary dynamics directly. Insofar as we want to make use of simple, intuitive, classifications of social behavior we need to resist drawing any deep or general evolutionary conclusions about those behaviors.

Yet classifying social interactions can structure our evolutionary investigations in an interesting way. The classification can help identify a puzzle or problem that more sophisticated studies can then address, but it is important to choose a classification scheme that is appropriate for the phenomenon under consideration. Hamilton's original classification scheme was aimed at a relatively simple type of altruistic social behavior. With more complex interactions, the original scheme lacks sufficient scope and resolution and hence cannot provide an appropriate guide to investigation.

Given the explanatory targets of the mutualistic approach, a classification scheme requires, at minimum, the resolution to capture the interactions described in Fig. 1. When we take seriously the possibilities revealed by the game-theoretic scheme, as we will argue below, it becomes clear that the mutualistic approach conflates different evolutionary scenarios, often treating Social Dilemma, Coordination, and Social Delight interactions as similar. In effect, the approach fails to identify the features of cooperative behavior with sufficient precision to offer an effective evolutionary explanation of the behavior.

Revisiting the mutualistic approach

Recall the shift that characterizes the mutualistic approach: the focus on mutually beneficial social interactions and the emphasis on partner choice for understanding the stability and complexity of prosocial behavior in human evolution. In what follows we will focus on the argument in Baumard et al. (2013) to show that dynamical considerations vastly complicate the role mutually beneficial behaviors may have played in human evolution. In particular, the appeal to partner choice complicates rather than simplifies the strategic context.

To start, note that mutually beneficial behaviors avoid one particular challenge that faces altruism: evolutionary stability. If some behavior, at or near fixation in a particular strategic context, counts as mutualistic then that behavior should be stable. Most alternative strategies won't be able to invade under usual circumstances.[10] Mutual benefit does not face the subversion problem that plagues altruism. If cooperation involves mutual benefit then it is easy to sustain, once it evolves. But how it evolves remains an open question—invasion may not be easy. Addressing the question of invasion involves discerning the relative advantages of cooperation in settings where it is rare and how, if rare, it could become more common. In the idealized framework of 2×2 symmetric games we need to know whether the strategic interaction involves Coordination or Social Delight. If it's the former, the evolution of cooperation requires crossing a fitness valley.

This point makes some statements made by Baumard et al. (2013) particularly puzzling. After introducing their distinction between partner control and partner choice (more on this below), they claim that "Mutually beneficial cooperation might in principle be stabilized either by partner control or by partner choice (or, obviously, some combination of both)" (61). This comment only makes sense if there is some threat of subversion, and therefore cooperation is altruistic and does not truly involve mutual benefit. Oddly, Baumard et al. (2013, fig. 1, 61) explicitly contrast mutualistic cooperation with altruistic cooperation, locating partner choice models squarely on the mutualistic side. The claim about the potential for partner choice to stabilize cooperation does not cohere with the claims that "mutualistic" cooperation is a new sort of behavior different from altruistic cooperation—the evolutionary obstacles that face the evolution of cooperative social behavior remain unchanged.

Contrast this with a similar but more nuanced claim in Sterelny (2012, 102):

> One important mechanism that reduces the impact of cheating is partner choice. ...in many circumstances, successful cooperation depends on identifying reliable partners while yourself being reliable, for many collaborations begin with mutual choice.

While Sterelny also sees partner choice as stabilizing cooperation, he avoids the problematic inconsistency with mutual benefit by treating partner choice as a

[10] Of course, better prosocial strategies can invade. And, if certain factors so conspire, such as non-random assortment, population structure, or the like, antisocial or selfish behavior can invade as well.

mechanism that helps against cheating. Cheating is only a problem in altruistic interactions. If cooperation is mutually beneficial, what incentive is there to cheat? Sterelny's comment occurs in the context of a discussion about the potential costs of defection and mechanisms for managing commitment to cooperative coalitions, thus explicitly recognizing the altruistic, Social Dilemma features of the strategic interaction, and how partner choice can influence that strategic context. It also makes clear the connection between cooperation and reputation. Tracking reputation is one way to manage indirect reciprocity and penalize defection which can further stabilize cooperation (Alexander 1987; Nowak and Sigmund 2005).

The mechanism of partner choice is proposed as an alternative to partner control. Partner control involves influencing the payoffs of partners to incentivize cooperation over repeated interactions. Direct reciprocity (reciprocal altruism) is a mechanism of partner control; exchanging altruistic acts over time makes prosocial cooperation mutually beneficial in the long term. The famous tit-for-tat strategy in the iterated Prisoner's Dilemma is also an example; cooperation is rewarded and defection is punished. The distinction between partner choice and partner control tracks whether individuals have the option to abandon individuals and seek out new partners rather than having to cope with the same individual over multiple interactions. This an important causal difference in how interactions occur across the population, but notice that the mechanisms of partner choice and control do not affect the underlying type of interaction.

Furthermore, the fitness functions in evolutionary game theory depend on the frequency of different types of interactions, and from this perspective both partner control and partner choice produce the same result: they transform the frequencies of different types of interactions taking place. In effect, control and choice both change the incentives for cooperation and defection, making altruistic behaviors look more like mutualistic ones. More precisely, under the right conditions, these mechanisms increase an individual's preference or fitness (all things considered) to answer cooperation with cooperation. If interaction partners can punish defection with defection in later interactions (control) or exclusion from cooperative coalitions (choice), this can increase the long term fitness of cooperation and decreases the long term fitness of defection.

Notice that this observation alone does not permit conclusions about anything more than the stability of cooperation. When cooperation is common, both partner choice and partner control help stabilize it, as many theorists have shown. But what about how cooperation fares when rare? Is the underlying interaction a case of Social Delight, Coordination, or something more complicated? With respect to the studies on indirect reciprocity, even relatively "simple" models can be very complex and difficult to analyze, often producing unanticipated results.[11] Without specifying the details of the interaction and how choice versus control may work in a particular population, the distinction fails to track any deep difference in how and why evolutionary dynamics tend to produce cooperation.

[11] Cooperation is stable in some settings and unstable in others. In settings where cooperation is stable, it may or may not evolve reliably. Errors in action, information transfer, and nuanced conditional behavior can all result in different evolutionary scenarios, some more favorable for cooperation than others. See Nowak and Sigmund (2005) for a survey of these results.

Let us look at partner choice in more detail. According to Baumard et al. (2013, 61), "in partner-choice models...a cooperator reacts to a partner's cheating by starting a new cooperative relationship with another hopefully more cooperative partner." How does the strategy of opting out affect the evolutionary dynamics? For this to create a case of mutual benefit, one (or more) conditions must be met: (1) there should be a minimal cost to finding new partners, otherwise opting out may be too costly to be an effective response to defection; (2) cooperators should be able to find other cooperators more effectively than defectors can, since defectors that exploit cooperators would rather pair with a cooperator than another defector; or (3) there should be a (relatively) small advantage gained by defection. More important, the possibility of parter choice does not address the question of how prosocial cooperation might invade. There needs to be very specific and effective mechanisms in place for cooperators to find one another in the population when rare. So, for example, the claim that "In a market of cooperative partners, the most cooperative individuals end up interacting with one another and enjoy more common good" (Baumard et al. 2013, 62) is true only under restrictive conditions. Either cooperators are common or they are extremely effective at seeking one another out. The structure of the "market of cooperative partners" is crucial to the evolutionary dynamics, and it is vastly underspecified. While, as Baumard et al. (2013, 77) say, "...the distribution of benefits in each interaction is constrained by the existence of outside opportunities determined by the market of potential partners," without any information on that market—whether potential cooperative partners are rare or common, how the costs of searching out a new partner compare to the costs of simply not interacting or interacting with a defector, and so on—we can make no claims about the evolutionary origin or stability of cooperation.

Eshel and Cavalli-Sforza (1982) provide an elegantly simple model of partner choice to investigate the role of positive assortment (correlated interaction) in the evolution of altruism. They show that while random interactions assumed in evolutionary game theory by Maynard Smith and Price (1973) lead to the extinction of altruism, positive assortment imposed by either population structure or "active choice of companions" favors altruism. To model active choice of companions they borrow a model of assortative mating and apply it to social behavior. Individuals have a certain number of meetings from which they choose an "encounter" (interaction partner). Unsurprisingly, they find that "a selectively cooperative type is always more successful whenever the number of meetings increases" (Eshel and Cavalli-Sforza 1982).

Even in this simple model, whether altruism evolves depends on a number of factors: the costs and benefits of cooperative behavior, the degree of assortment introduced by population structure, the costs associated with the population structure (maintenance of kin groups, dispersal range, etc.), and the number of expected meetings per individual. This study provides the details that are missing from the informal arguments about partner choice, and reveals some of the complications we must face when assessing the evolutionary significance of the marketplace of cooperators. If the conditions are just right then the informal arguments work. The trouble is that informal arguments often do not restrict their scope to when conditions are just right.

The model also points towards two factors not explicitly considered but would introduce evolutionary complications to partner choice: finite populations and fitness costs associated with increasing the number of meetings. Both are important factors if we want to understand the evolution of human cooperation. Introducing limitations to the number of possible interaction partners and costs to searching out new partners each disrupt the straightforward informal argument about partner choice favoring cooperation, even so-called mutualistic cooperation. Finite populations create a musical chairs element to encounters where individuals need to find a suitable interaction partner and fast, otherwise they may miss the opportunity to engage in cooperative social behavior entirely. Search costs discount the effectiveness of partner choice, perhaps making partner control a better alternative in some scenarios. Such factors could lead to the emergence of commitment devices, mechanisms that ensure an individual cannot opt out of tribes or coalitions, to stabilize cooperation through control rather than choice (Sterelny 2012). Partner choice does not solve all the evolutionary problems associated with cooperation.

Even in the best case scenario for cooperation, Social Delight, there are potential obstacles related to finite populations and the size of the cooperative payoffs (Forber and Smead 2014). Indeed, the devil is in the details for understanding the evolution of cooperation, and recognizing that some cooperative behavior involves mutual benefit does little to clarify the evolutionary picture. More generally, the variety of results and complexity of models suggest that it is not even clear that there is a coherent unified account of mutualistic cooperation to be had.

Additive fitness effects

Our conceptual criticism about the limitation of the standard scheme converges with a theoretical observation about Hamilton's inclusive fitness framework. The observation is that the framework assumes a kind of additivity to accurately predict evolutionary dynamics: the gains and losses of social behavior need to accumulate in a strictly additive way without interaction effects introduced by the recipient's behavior (Cavalli-Sforza and Feldman 1978; van Veelen 2009; Allen et al. 2013). The particular prediction of interest is Hamilton's rule: altruistic behavior should evolve when relatedness (effectively the correlation between types of behavior during interactions) is greater than the ratio of the cost of altruistic behavior to the benefit conferred on others (Hamilton 1964a). In one of the earliest studies to recognize this limitation to the inclusive fitness framework, Cavalli-Sforza and Feldman (1978) showed that if the costs and benefits are additive in the fitness functions for behavioral types, Hamilton's rule accurately predicts evolution. However, if the fitness effects are multiplicative (e.g., if the cost acts as a discount factor affecting the amount of benefit received), then Hamilton's rule no longer makes accurate predictions. Both the ratio of cost to benefit and the magnitude of the benefit affect whether altruistic social behavior evolves in the multiplicative case.

The point of these formal results can be put in an informal way by considering what inclusive fitness is supposed to represent. Since social behavior affects both the fitness of the focal individual and that of all her interaction partners (and vice versa), we need to account for these effects, if we are to predict the evolutionary fate of these behaviors. Hamilton developed the inclusive fitness framework to do precisely this. The most tractable scenario is when these effects are additive—we simply tally up the costs the actor pays out and the benefits that she receives. In this way I can determine the inclusive fitness for each individual in the population and the inclusive fitness for each behavioral type. Compare these fitnesses and we can make evolutionary predictions. But the world may be more complicated. An actor's behavior may interact with the recipient's behavior in ways that change the fitness effects of behavioral type across different interactions. For instance, there may be synergistic effects between cooperative behaviors, or perhaps there are diminishing marginal returns to altruism in larger groups. The existence of these effects can create dynamical scenarios whose evolutionary outcomes are not accurately predicted by Hamilton's rule (Queller 1985). We can no longer tally up costs and benefits; we need to change the magnitudes of the fitness effects based on the type and frequency of interactions the individuals experience.

The connection to our argument should now be clear: the strategic interactions modeled using game theory extend to a range of cases where the recipient's behavioral strategies change the fitness effects of the actor's behavior. We cannot compare the costs and benefits to a some baseline because those costs and benefits depend on the recipient's strategic repertoire. Contrast Table 4 with Table 5. We can obtain the 2×1 game as a special case of the more general game-theoretic framework assuming that the recipient's behavior has no effect on the fitness consequences of the actor's behavior. That is, we assume that the fitness effects of the actor's behavior are *additive* and do not change across different interaction types. That way we simply need to tally up the fitness effects of behaviors across a population of actors to make evolutionary predictions. With respect to altruism, Hamilton's approach considers only the strategic repertoire of the actor when determining the fitness gains and losses, and by summing these gains and losses across the population makes the prediction in Hamilton's rule: if cooperative actors tend to interact with other cooperative actors sufficiently often (i.e., when $r > c/b$) then altruism evolves. The inclusive fitness framework assumes that the costs and benefits of altruistic behavior remain the same whether the recipient behaves altruistically or selfishly. As soon as we consider strategic interactions, interactions when the recipient can respond to an actor in ways that change the fitness effects of social behavior, we need a richer framework to accurately represent the range of evolutionary scenarios.[12]

[12] This line of reasoning can generalize to combinations of alleles in diploid population genetic models and therefore identifies a limitation on optimality analyses of the evolution of recombination that assume additivity across alleles. See Feldman et al. (1997) for discussion.

The way forward

So far we have delivered a primarily critical message. If the mutualistic approach is to be abandoned, what is the alternative? Attending to the evolutionary dynamics and strategic contexts of social behavior, as we have done here, reveals a way forward. We should focus on the various models themselves by specifying details, analyzing evolutionary dynamics, and assessing external validity. The Eshel and Cavalli-Sforza (1982) model does this nicely for partner choice: the model focuses on altruism and explores how positive assortment, the cost associated with assortment, and ability to choose interaction partners from a number of possible encounters interact to facilitate the evolution of altruism. These are the sort of details that are not specified in the informal arguments found in Baumard et al. (2013). The traditional classification scheme simply does not have any substantive evolutionary implications in the settings of interest to the mutualistic approach. Nothing short of building and analyzing the evolutionary models will identify the possible outcomes.

That said, classifying the *interaction type*—Conflict, Coordination, Social Dilemma, or Social Delight in the 2 × 2 case—can be helpful in a more refined way. Classification identifies the evolutionary challenges that face particular social behaviors. In the case of Social Dilemmas, prosocial cooperation needs to overcome the traditional problem of subversion. In the case of Coordination, cooperation is stable once it evolves, so subversion is no longer an evolutionary obstacle, but it is risky when rare, posing a problem for invasion. In the case of mutual benefit, neither stability nor invasion are particularly problematic. Once we determine the interaction type, we can use evolutionary models to explore whether and under what conditions different sorts of *interaction structure*—positive assortment, conditional strategies, repeated interactions, spatial structure, and so on—can produce the evolution of cooperation. To put it another way, the interaction type sets the explanatory target and the interaction structure provides part of the evolutionary explanation.[13]

Our critical comments also motivate two important positive projects for exploring the evolution of social behavior. The first involves the coevolution of social behavior and the interaction structure. Skyrms (2004) presents model along these lines where social behavior co-evolves with dynamic association networks. He finds that it is possible for cooperation to evolve in Social Dilemma and Coordination games. However, in both cases, it is also possible for defection to evolve and the prospects for cooperation depend on the specific conditions and parameters of the model. Lehmann et al. (2009) provide an another example. They investigate the coevolution of conditional helping and harming behavior with neutral markers using two-locus population genetics, allowing the interaction

[13] It is important to note that this contrast between interaction type and interaction structure may collapse at higher levels of abstraction. For example, a repeated Prisoner's Dilemma with a restricted strategy space has the same strategic dynamics as a one-shot Stag Hunt (Skyrms 2004). This shows that there are cases where a given interaction type and interaction structure can be abstractly represented with another interaction type with a different interaction structure. Thus, the distinction between interaction type and interaction structure should be treated as a methodological tool rather than a hard distinction.

structure to emerge through the evolutionary process. The study finds that both marker-based conditional helping and harming are potential evolutionary outcomes, but that, perhaps surprisingly, "if everything else is equal, marker-based conditional harming is often more likely to evolve than marker-based conditional helping" (Lehmann et al. 2009, 2896). It is a good bet that these results complicate the informal arguments often made about mutualistic cooperation.

The second positive project involves the conditions under which social behaviors can generate fluid strategic contexts. A common theme from the mutualistic approach is that the key to unlocking the evolution of human cooperation is, in some sense, to change the underlying game. In terms of the idealized framework for 2×2 games, Social Dilemmas can be changed into Coordination games, or perhaps even into social delights in the right circumstances. Allowing the game or interaction type to co-evolve with behavioral strategies unlocks a novel space of dynamical scenarios (Hashimoto and Kumagi 2003; Worden and Levin 2007; Smead 2014). Alternatively, individuals may be faced with two or more interaction types but not know which game they are playing at a particular time. In this case behaviors that are adaptive in one setting might be co-opted in other settings; playing fair, for instance, may help deal with the uncertainty about interaction types even though it might not be the best response in particular games (Harsanyi 1967; Bednar and Page 2007; Zollman 2008). Understanding how effective cooperation in some contexts might lead to prosocial results in other settings is an essential part of a complete account of the evolution of cooperation.

In short, the evolution of social behavior is strikingly complicated. A classification scheme can help identify the evolutionary obstacles that face different sorts of behavior. But such a scheme requires the resolution to specify the relevant obstacles. Without that resolution, investigations fail to engage the complete constellation of factors that affect the evolutionary dynamics of social behavior, especially cooperation.

Acknowledgments Thanks to Elliott Sober, Marty Barrett, Malcolm Forster, two anonymous referees, and the audience at POBAM 2014 for valuable feedback and discussion.

References

Alexander R (1987) The biology of moral systems. Aldine Transaction, Chicago

Allen B, Nowak MA, Wilson EO (2013) Limitations of inclusive fitness. Proc Natl Acad Sci 110:20135–20139

Ariew A, Lewontin RC (2004) The confusions of fitness. Br J Philos Sci 55:347–363

Baumard N, André JB, Sperber D (2013) A mutualistic approach to morality: the evolution of fairness by partner choice. Behav Brain Sci 36:59–122

Beatty J (1992) Fitness: theoretical contexts. In: Keller EF, Lloyd EA (eds) Keywords in evolutionary biology. Harvard University Press, Cambridge, pp 115–119

Beatty J, Finsen S (1989) Rethinking the propensity interpretation: a peek inside pandora's box. In: Ruse M (ed) What philosophy of biology is. Kluwer Academic Publishers, Dordrecht, pp 17–30

Bednar J, Page S (2007) Can game(s) theory explain culture? The emergence of cultural behavior within multiple games. Ration Soc 19:65–97

Bomze IM (1983) Lotka–Volterra equation and replicator dynamics: a two-dimensional classification. Biol Cybern 48:201–211

Cavalli-Sforza LL, Feldman MW (1978) Darwinian selection and "altruism". Theor Popul Biol 14:268–280
Eshel I, Cavalli-Sforza LL (1982) Assortment of encounters and the evolution of cooperativeness. Proc Natl Acad Sci 79:1331–1335
Feldman MW, Otto SP, Christiansen FB (1997) Population genetic perspectives on the evolution of recombination. Annu Rev Genet 30:261–295
Forber P, Smead R (2014) An evolutionary paradox for prosocial behavior. J Philos 111:151–166
Frank SA (1998) Foundations of social evolution. Princeton University Press, Princeton
Hamilton WD (1964a) The genetical evolution of social behaviour, I. J Theor Biol 7:1–16
Hamilton WD (1964b) The genetical evolution of social behaviour, II. J Theor Biol 7:17–52
Hamilton WD (1970) Selfish and spiteful behavior in an evolutionary model. Nature 228:1218–1220
Harsanyi J (1967) Games with incomplete information played by "Bayesian" players. Manag Sci 14(3):159–182
Hashimoto T, Kumagi Y (2003) Meta-evolutionary game dynamics for mathematical modeling of rules dynamics. In: Banzhaf B, Christaller T, Ziegler J (eds) Advances in artificial life. Springer, Berlin, pp 107–117
Hauert C, Michor F, Nowak MA, Doebeli M (2006) Synergy and discounting of cooperation in social dilemmas. J Theor Biol 239(2):195–202
Lehmann L, Bargum K, Reuter M (2006) An evolutionary analysis of the relationship between spite and altruism. J Evol Biol 19:1507–1516
Lehmann L, Feldman MW, Rousset F (2009) On the evolution of harming and recognition in finite panmictic and infinite structured populations. Evolution 63:2896–2913
Maynard Smith J, Price GR (1973) The logic of animal conflict. Nature 246:15–18
Nowak MA, Sigmund K (2005) Evolution of indirect reciprocity. Nature 437:1291–1298
Okasha S (2007) Evolution and the levels of selection. Oxford University Press, Oxford
Price GR (1970) Selection and covariance. Nature 277:520–521
Queller D (1985) Kinship, reciprocity and synergism in the evolution of social behavior. Nature 318:366–367
Skyrms B (1996) Evolution of the social contract. Cambridge University Press, Cambridge
Skyrms B (2004) The stag hunt and the evolution of social structure. Cambridge University Press, Cambridge
Smead R (2014) Evolving games and the social contract. In: Youngman PA, Hadzikadic M (eds) Complexity and the human experience. Pan Stanford, Singapore, pp 61–80
Smead R, Forber P (2013) The evolutionary dynamics of spite in finite populations. Evolution 67(3):698–707
Sober E (2001) The two faces of fitness. In: Singh RS, Krimbas CB, Paul DP, Beatty J (eds) Thinking about evolution. Cambridge University Press, Cambridge
Sober E, Wilson DS (1998) Unto others. Harvard University Press, Cambridge
Sterelny K (2012) The evolved apprentice: how evolution made humans unique. MIT Press, Cambridge
Sterelny K, Joyce R, Calcott B, Fraser B (2013) Cooperation and its evolution. MIT Press, Cambridge
Taylor HM, Gourley RS, Lawrence CE, Kaplan RS (1974) Natural selection of life history attributes: an analytical approach. Theor Popul Biol 5:104–122
Tomasello M, Vaish A (2013) Origins of human cooperation and morality. Annu Rev Psychol 64:231–255
Tomasello M, Melis AP, Tennie C, Wyman E, Herrmann E (2012) Two key steps in the evolution of human cooperation: the interdependence hypothesis. Curr Anthropol 53:673–692
van Veelen M (2009) Group selection, kin selection, altruism and cooperation: when inclusive fitness is right and when it can be wrong. J Theor Biol 259:589–600
Weibull JW (1995) Evolutionary game theory. MIT Press, Cambridge
West SA, Gardner A (2010) Altruism, spite, and greenbeards. Science 327:1341–1344
West SA, Griffin AS, Gardner A (2007) Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. J Evol Biol 20:415–432
Worden L, Levin SA (2007) Evolutionary escape from the prisoner's dilemma. J Theor Biol 245:411–411
Zollman KJS (2008) Explaining fairness in complex environments. Polit Philos Econ 7(1):81–98