Title
Indirect reciprocity and the evolution of “moral signals”

Journal
Biology & Philosophy, 25(1)

ISSN
1572-8404

Author
Smead, Rory

Publication Date
2010

DOI
10.1007/s10539-009-9175-9

Peer reviewed

Biol Philos (2010) 25:33–51
DOI 10.1007/s10539-009-9175-9

Indirect reciprocity and the evolution of "moral signals"

Rory Smead

Received: 2 November 2008 / Accepted: 23 June 2009 / Published online: 9 July 2009

Abstract Signals regarding the behavior of others are an essential element of human moral systems and there are important evolutionary connections between language and large-scale cooperation. In particular, social communication may be required for the reputation tracking needed to stabilize indirect reciprocity. Additionally, scholars have suggested that the benefits of indirect reciprocity may have been important for the evolution of language and that social signals may have coevolved with large-scale cooperation. This paper investigates the possibility of such a coevolution. Using the tools of evolutionary game theory, we present a model that incorporates primitive "moral signaling" into a simple setting of indirect reciprocity. This model reveals some potential difficulties for the evolution of "moral signals." We find that it is possible for "moral signals" to evolve alongside indirect reciprocity, but without some external pressure aiding the evolution of a signaling system, such a coevolution is unlikely.

Keywords Cooperation · Evolution · Indirect reciprocity · Language · Morality · Moral signals · Reputation · Signaling

Introduction

Language is an essential component in human cooperative social systems, allowing fast and efficient information exchange in addition to facilitating social monitoring and reputation tracking. Did the benefits of social cooperation provide the selective pressure that caused the evolution of moral language? And, could such a language have evolved alongside social cooperation? To take a first step in answering these

R. Smead (✉)
Department of Logic and Philosophy of Science, School of Social Sciences,
University of California, Irvine, 3151 Social Science Plaza A, Irvine, CA 92697, USA
e-mail: rsmead@uci.edu

broad questions, we will examine the prospects for the evolution of primitive "moral signals" in the context of cooperation through indirect reciprocity.

The importance of reciprocity to the evolution of altruism was suggested by Trivers (1971) and direct reciprocity has since received a considerable amount of attention from scholars including the well-known computer simulations of Axelrod (1984). Later, Alexander (1987) argued that *indirect reciprocity* also plays an extremely important role in human moral systems. The idea is that an agent will behave altruistically toward one person so that she will benefit from altruistic acts of others: if you scratch my back, someone else will scratch your back.

Recently, there has been enormous progress made on understanding how cooperation through indirect reciprocity could have evolved. Most notably, Nowak and Sigmund (1998b) have provided a simple model where agents may condition their behavior on the "image" (or reputation) of other individuals. The model includes a discriminating strategy, which selectively punishes those with poor images. Such a discriminating strategy provides a sort of community enforcement against defectors and thus enables the evolution of large-scale cooperation.[1]

However, these basic models do not provide an explicit image-tracking mechanism beyond direct observation. Nowak and Sigmund (1998b, 2005) suggest informally that language may provide such a mechanism and that there may have been a coevolution of social communication and indirect reciprocity. "The evolution of human language as a means of information transfer has certainly helped in the emergence of cooperation based on indirect reciprocity" (Nowak and Sigmund 1998b). And, "Indirect reciprocity requires information storage and transfer as well as strategic thinking and has a pivotal role in the evolution of collaboration and communication" (Nowak and Sigmund 2005). This same sentiment has been echoed by Sterelny (2003): "Language is superbly adapted for social monitoring." These ideas are also related to Joyce's characterization of the "gossip hypothesis" for the evolution of language: "…human linguistic faculties were selected for in order to serve reciprocal exchanges when the groups got large. A language of gossip is a language of reciprocity" (Joyce 2007).[2] The evolution of language generally is too large a topic for this paper.

Instead, we will focus on the feasibility of a simultaneous evolution of primitive social communication and indirect reciprocity. Classifying an individual, or her actions, through language as "good" or "bad" would provide the reputation tracking needed for indirect reciprocity to arise. And, given the sentiments mentioned above, it is intuitively plausible that there could have been a coevolution of indirect reciprocity and some social signaling system. Along these lines, Harms and Skyrms (2008) have informally suggested that incorporating a simple signaling game into the models of Nowak and Sigmund may provide the beginnings of an

[1] Kandori (1992) was of the first game theorists to study community enforcement.

[2] The gossip hypothesis is motivated by studies of Aiello and Dunbar (1993), Dunbar (1993, 1996). It is arguably far more nuanced than this quote suggests and is a very interesting and controversial topic that demands a more detailed analysis than can be given in this paper. Here, we will restrict ourselves to the general idea that the benefit of large-scale cooperation was a major selective force in the evolution of language.

evolutionary account of "moral signals."[3] The idea is that a primitive social signaling system, which tracks the behavior of individuals and facilitates enforcement of cooperative norms, may be able to evolve in a setting of indirect reciprocity. Such a signaling system would allow a population to reap the benefits of indirect reciprocity and may be representative of a primitive moral language.

It is surely true that language plays an important role in indirect reciprocity. Even so, this does not necessarily mean that language evolved in this context or that it evolved *because of* the benefits gained by cooperation. One aim of this paper is to gain some theoretical insight into the idea that *the benefit of large-scale cooperation (by indirect reciprocity) was a major selective force in the evolution of language (as social communication)*. By doing so, we will be able to form a limited assessment of this claim's plausibility.

To accomplish this, we will pursue the suggestion of Harms and Skyrms (2008) by modeling image scores as simple social signals between players and determine the prospects for the coevolution of social signaling and indirect reciprocity.[4] This paper will advance beyond the current models of indirect reciprocity by explicitly modeling the evolution signaling as a mechanism for reputation tracking. In the models studied here, image scores will be determined by the signals of others and the signaling strategies are also subject to evolution. This will provide a first step toward an understanding of the evolution of "moral signals" and allow us to theorize more precisely about the effect of cooperative benefits on the evolution of language. Ultimately, we will see that there are ways cooperation and moral signals can coevolve. However, without some additional exogenous pressure on the signals in the simple settings we consider, such a coevolution is unlikely. These results highlight the difficulties faced in providing theoretical foundations for the claim that the benefits of indirect reciprocity were a driving force in the evolution of social communication.

The evolution of indirect reciprocity

Indirect reciprocity is when an individual A receives aid from another individual B because A previously helped individual C.[5] The existence of this sort of behavior is somewhat of a puzzle for evolutionary theorists. Direct reciprocity is less of a puzzle, because the aid is coming from the person who received it and can be withheld as punishment for previous uncooperative behavior. C may have an interest in reciprocating with A because of benefit from future interactions, but why would B have any interest in facilitating cooperation between A and C by providing

[3] The term "moral signals" is drawn from Harms and Skyrms (2008). The focal point of their discussion, and the starting point for the model explored in this paper, is "cheap-talk" or cost-free signaling rather than costly signaling. These games are explained in more detail in Section "Moral signals".

[4] Including a signaling component in other games has produced some interesting results; see Skyrms (2002, 2004) and Zollman (2005).

[5] There are two varieties of indirect reciprocity. "Upstream" indirect reciprocity is when A lends aid to C and because of that, B helps C. "Downstream" indirect reciprocity is when A lends aid to C and because of that, B helps A. Both of these behaviors are seen in humans, but the focus of this paper is downstream indirect reciprocity.

punishment or reward (at a cost to herself)? The answer to this question may lie in the nature of the community and community enforcement. *B* will want to play her role in this exchange so that she will elicit future cooperation from others. Cooperation is sustained not by individual enforcement but by community enforcement where the population behaves according a social norm of helping those who give help.

The key to evolving cooperation through indirect reciprocity is an individual's image or reputation. Without some way for agents in the population to distinguish between cooperators and defectors, there would be no way for indirect reciprocity to work. Reputation provides a method of identifying cooperators. Nowak and Sigmund (1998a, b) show that a simple binary image score which tracks only an agent's most recent action is sufficient to allow for evolution of cooperation by indirect reciprocity.

A simple model of image scoring

The model of Nowak and Sigmund (1998a, b) is simplistic relative to the actual complexity of indirect reciprocity in humans. However, it reveals some important and interesting aspects of indirect reciprocity and its evolution. We can imagine a large population of players randomly paired a fixed number of times in the role of donor or recipient. The donor may choose to incur a cost c to give a benefit b to the recipient with $b > c$. Each player has an "image score" based on past actions. In the simplest setting, there are only two image scores: "good" (or 0) which means the agent donated on her last action and "bad" (or 1) which means the agent refused donation on her last action.[6] There are three strategies: always donate (cooperate), never donate (defect), and the image scoring strategy of donate if and only if the other guy is "good" (discriminate).

A fitness value is assigned to each strategy based on the accumulated payoffs of the interactions. Differences in fitness cause the proportion of each strategy in the population to increase or decrease according to the replicator dynamics (Taylor and Jonker 1978). These dynamics can be interpreted in either a biological context (as differential reproduction) or in a cultural context (as differential imitation or learning of strategies). On this model, cooperative states can evolve provided the initial proportion of discriminators is high enough to drive the defectors to extinction. The resulting cooperative populations are neutrally stable mixes of discriminators and unconditional cooperators. Since these strategies are wholly cooperative in the absence of defectors, each receives an identical payoff in these states. Although this result does not necessitate cooperation, it does show that image scoring strategies can stabilize it. Figure 1 provides a basic picture of the dynamics of indirect reciprocity in the simple image scoring model.[7] This dynamical picture closely resembles the situation with the evolution of *direct* reciprocity.

[6] Nowak and Sigmund (1998b) also present a model with many image scores ranging from + 5 to −5, and find that this setting is also conducive to the evolution of indirect reciprocity.

[7] This figure was generated by simulations using the discrete-time version of the replicator dynamic (Weibull 1995).

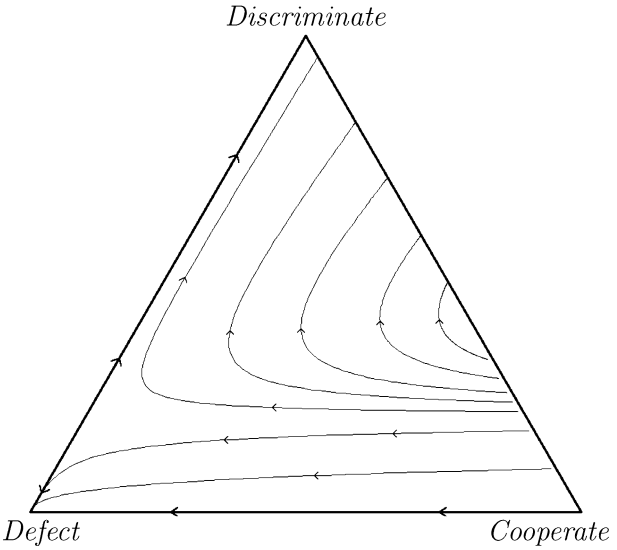

Fig. 1 The dynamics of indirect reciprocity

These results are promising, but many subtle problems arise when complexities are introduced. For instance, if there are errors in perception or action (i.e. the image scores of other players cannot be reliably accessed or defection sometimes occurs unwillingly), cooperation can be hindered and destabilized (Panchanathan and Boyd 2003; Nowak and Sigmund 2005; Brandt and Sigmund 2005a).[8] Additionally, Leimar and Hammerstein (2001) have argued that the image scoring strategy is "paradoxical": why should discriminators punish others (by defecting) if they pay the cost of having a damaged reputation themselves? To avoid these problems, many have turned to the investigation of "standing strategies." Standing strategies attend to both the action performed and the circumstances in which it was performed. This allows discriminators to treat defecting on defectors as a good thing rather than a bad thing.[9] Many models have been provided which show the success of standing strategies in stabilizing cooperation.

For our purposes here, we will focus only on the basic binary image scoring model from Nowak and Sigmund (1998a). The primary motivation for this focus is simplicity: models which include both second-order standing strategies and signaling will be enormously complex, making any results very difficult to interpret properly. There are also empirical results that emphasize the importance of image scoring. For instance, Milinski et al. (2001) examined indirect reciprocity

[8] Fishman (2003) also discusses errors in action. He shows that a certain degree of "forgiveness" in the discriminating strategy can stabilize cooperation, which becomes undermined with "complete fidelity" in image scoring. Additionally, Brandt and Sigmund (2005b) show that if the probability of errors in perception decreases over an individual's lifetime, cooperation can be stabilized.

[9] The notion of a standing strategy was first introduced by Sugden (1986). For more work on standing strategies see Panchanathan and Boyd (2003), Ohtsuki and Iwasa (2004), Leimar and Hammerstein (2001), Nowak and Sigmund (2005).

experimentally, finding that subjects tended to behave in accordance with image scoring strategies.

Setting aside standing strategies, there are two questions in the present context that need to be answered regarding the simple image scoring model. Is it possible to observe a coevolution of indirect reciprocity and the social signals that will serve as a basic image-tracking mechanism? And, if such a coevolution is possible, what implications does this have for the evolution of primitive "moral signals"? To answer these questions, we will introduce a simple signaling game into the model: agents will send signals regarding the most recent behavior of others.

"Moral signals"

Harms and Skyrms (2008) describe the research project surrounding the evolution of moral norms as having three connected explanatory targets: behavior in accordance with norms, enforcement of norms by punishment, and "moral signals." The first of two of these have received the vast majority of attention from scholars. The third has been given very little formal treatment and the aim here is to advance our understanding of this area by working with a simple model that connects a social signaling game and the models of indirect reciprocity. It should be noted that providing a full evolutionary account of moral language would be a massive undertaking well beyond the scope of this paper; human moral language, and its relationship to moral judgments and actions is far too complicated a phenomenon for simple game theoretic models to account for. However, given the recent work on indirect reciprocity and the evolution of communication in simple signaling games we are in a position to take a first step in advancing our understanding of this particular feature of human moral systems.

Moral language and moral signals enable, among many things, (i) the transmission of information about social behavior and (ii) the facilitation of enforcement of social norms. Here we use the term "moral signals" (with quotes) to designate this specific two-part role. In the models presented here, it will be possible for the population to use the signals in a way that fits with each of these parts. It is in this limited sense that these models will allow us to analyze the feasibility of "moral signals" evolving alongside indirect reciprocity. Determining if and how "moral signals" can evolve in this setting will be an important part of two different explanatory projects. The first is explaining the evolution of moral norms. The second is explaining the evolution of an image-tracking mechanism for indirect reciprocity.

To model the evolution of "moral signals" a simple sender–receiver game (Lewis 1969; Skyrms 1996) will be introduced into a binary-image model of indirect reciprocity. The goal of the binary Lewis sender–receiver game is to match one of two states of nature (observed by one individual) with a corresponding correct action (taken by another individual) by means of sending one of two signals. The game proceeds as follows. Nature chooses a state (*left* with probability q or *right* with probability $1 - q$). The sender observes the state and can send one of two messages (A or B) to the receiver. The receiver observes the message (but not the

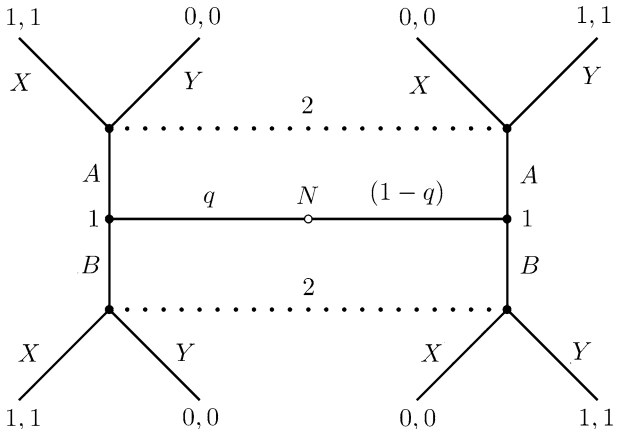

Fig. 2 Sender–receiver game in extensive form

state) and can choose one of two actions (X or Y). Each player is paid off 1 if the act matches the state (X for left, Y for right) and 0 otherwise. Figure 2 shows the sender–receiver game in extensive form. Here, a signaling system is a set of strategies that guarantees maximum payoff: one that is perfectly communicative. There are two signaling systems in this game, and since there is no natural salience to the signals, these solutions are "conventional." In the evolutionary setting, where a population is playing a signaling game and different strategies reproduce according to relative success, selection will often lead to one of these conventional signaling systems.[10]

Once an evolving population reaches a signaling system equilibrium, the signals used are primitive in that they are neither clearly indicative nor clearly imperative. In such populations, signals are both perfectly correlated with the state of the world and with the action taken in response. Thus, we can interpret a message in a signaling system as either "Do X" or "the state is left" equivalently: they have *primitive content* (Harms 2004) or are "Pushmi-Pullyu" because they "have both a descriptive and directive function" (Millikan 2005). Harms (2000) argues that there is good reason to think that basic "moral signals" are similar in this respect. Intuitively, "moral signals" should both carry information about others and facilitate certain behaviors toward them.

In the model presented below, a signaling game structure will be imbedded in the setting of indirect reciprocity and agents will communicate about other individuals rather than about the state of the world. Individual A is able to send messages regarding their past interaction partner B to B's future interaction partner C.

Very few models have attempted to explicitly capture the evolution of such social communication. One such study is done by Nakamaru and Kawata (2004), who examine a model of indirect reciprocity that includes a form of communication they call "rumors." Their model includes strategies for starting and spreading rumors

[10] For more recent work on evolution in simple signaling games of this sort see Skyrms (1996, 2002, 2004), Zollman (2005), Huttegger (2007a, b), Barrett (2008).

about oneself or others. They argue that rumors will evolve if they can be used to detect and punish defectors. Additionally, which strategies are able to repel cheaters depends on the rate of communication in the population. The model presented in this paper will differ from that of Nakamaru and Kawata (2004) in several ways. In particular, Nakamaru and Kawata examine a more restricted set of signaling strategies in addition to including complexities such as variation on the rate of rumor communication, rumors about oneself, and the possibility of repeating play with a single individual. The model below will avoid these complexities and focus on a more direct form of communication without placing restrictions on the signaling strategies. Not restricting the strategy space, as we will see in the section "Including a payoff to signaling", allows for different methods of "moral signaling" to evolve that are similar to the different conventions of the Lewis sender receiver-games.

The model

The base game for the model presented here will be the Prisoner's Dilemma.[11] The payoff matrix shown below will provide a payoff function π, which will map a pair of strategies to a payoff.[12] In this model, members of an infinite, randomly mixing population will play a fixed number N of 1-shot games, each against a new opponent. In each interaction, players choose to either cooperate (c) or defect (d). Then, both players will have the opportunity to send a signal to the next player who interacts with her current opponent (either 0 or 1). We will begin by investigating the case where there is no direct benefit or cost associated with sending particular signals.

The Prisoner's Dilemma

	c	d
c	10,10	0,11
d	11,0	1,1

A strategy here is an ordered pair of functions (R,S). R maps signals to actions: $\{0,1\} \rightarrow \{c,d\}$ dictating how the player will act in the stage game. S maps actions to signals: $\{c,d\} \rightarrow \{0,1\}$ dictating what the player will signal about her previous opponent.[13] Let *Strat* represent the set of all strategies. There are 16 strategies in all, which can be represented as ordered quadruples such as (c,d,0,1) meaning "do c if

[11] The payoffs here can be varied while preserving ordinal ranking and produce similar results. For the simulations, these specific payoffs were chosen because they capture the general structure of cooperative or altruistic games while offering a substantial benefit to settling on the cooperative outcome.

[12] The game used by Nowak and Sigmund (1998a, b) is similar to a Prisoner's Dilemma if we take the expected payoffs for a 50% chance of being a donor and a 50% chance of being a receiver for each interaction.

[13] The signaling component of a strategy will be referred to as "separating" if it sends different signals for different actions.

opponent has a 0 image, do d otherwise, send 0 if opponent plays c and send 1 otherwise." Additionally, each player p in round n will have an image score $k_p^n \in\{0,1\}$ that is determined by the signal of their previous opponent q. When players interact, each responds to the image of her opponent and receives a payoff according to the game matrix above. Then, after the interaction, each player gives their opponent a new image before going on to new interactions. For simplicity we will assume that all players have an initial image score of 0.[14] Note that the signal q sends about p depends on the strategy of q and the action of p (which is a reaction to q's image). Thus, $k_p^{n+1} = S_q(R_p(k_q^n))$ and the payoff to p for an interaction with q on round n is then $\pi(R_p(k_q^n), R_q(k_p^n))$.[15]

The payoff of one strategy type i against another j cannot be calculated for a particular interaction without knowing the image scores of each individual, which depend on the other types in the population. Consequently, each round we calculate the probability that a given type has a particular image based on expected interactions in the previous round of play. We then calculate the expected utility of strategy i against strategy j under all possible image score combinations by weighting the payoffs with the appropriate probabilities. Let $u^n(i,j)$ denote the expected utility of using i against j in round n.

Individuals will play a fixed strategy and we are interested in the evolution of the strategy frequencies in a population. The fitness of a strategy type i in round n is based on the expected payoffs against the population in that round:

$$f_i^n(X) = \sum_{j \in Strat} u^n(i,j)x_j$$

where x_j represents the proportion of strategy type j in the population and $X = (x_1 \ldots x_{16})$ represents the current distribution of all strategy types in the population. The total fitness for the game $f_i(X)$ is the sum of the fitness from all rounds of play: $f_i(X) = \sum_n f_i^n(X)$. Calculating total fitness involves calculating rounds one by one while tracking the image scores generated from previous rounds.

In simulations, evolution occurs according to the discrete time replicator dynamics. The idea is that the strategies with higher fitness increase at the expense of those with lower fitness:

$$x_i' = x_i\left(\frac{f_i(X)}{\theta(X)}\right)$$

where θ is the average fitness of the population and x_i' is the frequency of type i at the next time step.[16]

[14] This adds a natural asymmetry into the model where some strategies start out cooperative (without information), others do not. For instance, since all individuals begin with an image of 0, the strategy $(c,d,0,1)$ cooperates in the absence of information and $(d,c,1,0)$ defects. As will be discussed in the section "Including a payoff to signaling", this can lead to the evolution of different possible "moral systems."

[15] It is easy to see that the effect of the signal from a player's first opponent may have ongoing effects several rounds later; to determine the outcome of a particular interaction, the histories of the players are needed. Because of this complexity, the game being played involves many individuals and is too complex to be represented in a simple form.

[16] Interested readers may contact the author for further details on the computer simulations.

Interpretation of the model

It is important to briefly discuss the interpretation of signals in this model. The signals "0" and "1" do not have any predetermined meaning such as "good" or "bad." However, in certain populations these signals can come to have *primitive content* in the sense described above. For instance, in a population of (c,d,0,1) the signal of "1" regarding an individual A could be loosely translated as "A defected; defect on A." This same population is highly cooperative, able transmit information about individuals, and able to use that information to punish (via defection) any non-cooperative behavior. Thus, the population would be using "moral signals" in the limited sense we are working with.[17]

Admittedly, this is an oversimplified model of human moral systems. Real societies differ in several ways; they have more complex interactions, more complex communication, and nuanced differences between reputation and moral standing. Despite these idealizations, however, we can use these models to get a sense of what factors may (or may not) be involved in the evolution of primitive social signals. Furthermore, first attempts at modeling social phenomena should focus on simple settings so as to gain a deeper understanding of the underlying dynamics before further complexities are introduced.

Within this model and various extensions considered below, we can determine how likely it is for indirect reciprocity to evolve alongside a system of "moral signals." We can also provide an account of the important factors in such a coevolution. This is the topic of the next two sections. If it turns out that "moral signals" evolve only very rarely or that their evolution requires substantial pressure from outside the domain of indirect reciprocity, then this may (in the absence of further modeling) give us a reason to doubt the importance of indirect reciprocity in the evolution of primitive moral language.

Signaling as reputation tracking?

In order to understand the results of the model, it is important to set a benchmark for comparison. Here, we will use the results of the simple binary-image models of indirect reciprocity for comparison. If we include only strategies that agree on signaling behavior and correspond to the cooperator (c,c,0,1), the discriminator (c,d,0,1), and the defector (d,d,0,1), Fig. 3 shows we can achieve similar results to Nowak and Sigmund (1998a) (compare to Fig. 1). On the model presented above, with $N = 5$ and the payoff structure given, simulations show that 79% of populations reach some sort of cooperative equilibrium which is a mixture between discriminators and cooperators. A population will be counted as cooperative if the majority of the interactions are cooperative. We can now compare this benchmark case to cases that include a variety of additional signaling strategies. This allows us to judge the

[17] One may object that there is no need to interpret these signals *normatively*. Even so, the behavioral and social role that "moral signals" are playing in these populations is at least one important aspect of human moral systems worthy of philosophical investigation.

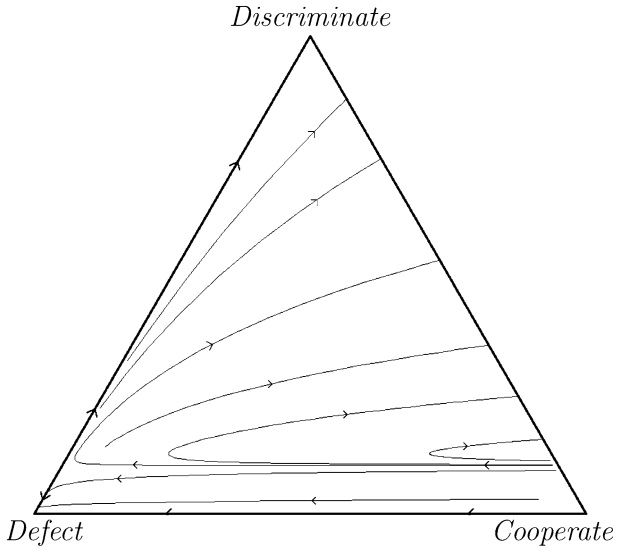

Fig. 3 The dynamics of indirect reciprocity

impact that a coevolving signaling system has on the evolutionary prospects for cooperation.

With the introduction of a signaling component, there are two new complications that could cause problems for the evolution of indirect reciprocity. The first complication is strategies that have an "inverted" view of what the signals mean relative to the population. Consider (d,c,0,1) relative to our benchmark case. This strategy has the same signaling behavior as the other strategies in the population, but defects on cooperative players and cooperates with the defectors. The second complication is strategies that signal in ways that would disrupt reputation tracking, such as always sending one signal or sending the opposite signals relative to the rest of the population. In general, the worry is that having an indeterminate signaling system serving as the mechanism for reputation tracking will, in effect, introduce errors in perception of image scores and destabilize cooperation as seen by Panchanathan and Boyd (2003).

Simulation results

These simulations were done using the discrete time replicator dynamics, a random initial distribution of strategy types in the population, the payoffs presented above, and $N = 5$ rounds of play. Table 1 shows the tendencies for cooperative outcomes when we include different sets of strategies. The first thing to note is that when we include the strategy which inverts the behavior of the traditional discriminators (d,c,0,1), even without including strategies that differ in their signaling behavior, there is a dramatic effect on the prospects for reaching cooperative states. Although this strategy does not survive in any resulting states, its presence certainly changes the evolutionary picture, hindering the evolution of cooperation by indirect

Table 1 Simulation results

Strategies	% Cooperative	Efficient?
(*c*,*c*,0,1), (*c*,*d*,0,1), (*d*,*d*,0,1)	79	Yes
(*c*,*c*,0,1), (*c*,*d*,0,1), (*d*,*d*,0,1), (*d*,*c*,0,1)	39	Yes
(*c*,*c*,0,1), (*c*,*d*,0,1), (*d*,*d*,1,0)	48	Yes
All strategies	<1	No
(*c*,*c*,1,1), (*c*,*d*,0,1), (*d*,*d*,0,1)	0	No

reciprocity. Simulations including the three traditional strategies and (*d*,*c*,0,1) result in cooperative states only 39% of the time.

Turning to the other possible complication—including different signaling strategies within the population—reveals that the prospects for a coevolution of signaling and indirect reciprocity are very grim. If we suppose that just the defectors signal in a way to disrupt the reputation tracking (*d*,*d*,1,0) then cooperation is hurt but still evolves a large portion of the time (48%). However, when all the strategies are included, the results for cooperation by indirect reciprocity are disastrous. Only 5 populations in 1,000 simulations reached cooperative states. To make things even worse, none of these "cooperative" states were efficient (average payoff of 10 per interaction), meaning that there was occasional defection.

The difficulties for cooperative populations are not entirely due to defecting individuals. Cooperative individuals who disrupt the image tracking signals can destabilize indirect reciprocity. For instance, one strategy that is particularly troublesome relative to the strategies in our benchmark case is the "Two-Faced" strategy which always cooperates, but always signals "1" (*c*,*c*,1,1). In a population of discriminators of type (*c*,*d*,0,1), a player using the Two-Faced strategy undermines the reputation of others without ever harming her own image. This allows the Two-Faced strategy to invade by causing the discriminators to begin defecting on each other and thereby paving the way for defectors to take over. Figure 4 shows the evolutionary dynamics on the face of the simplex which has only Two-Faced cooperators (*c*,*c*,1,1), signaling discriminators (*c*,*d*,0,1) and signaling defectors (*d*,*d*,0,1). Simulations involving only these three strategies result in uncooperative populations every time. This reveals that not only does the presence of communication errors undermine cooperation, as shown by the perception errors examined by Panchanathan and Boyd (2003), but that the situation is much worse: *an error-causing strategy can invade a population of error-free discriminators*.

With all possible strategies present, the only populations that reached cooperative states were those that began with a very high proportion of strategies labeling cooperation with "0" and defection with "1". In this simple model, without some way to police signals and enforce conformity to a separating signaling system, the prospects for the coevolution of reputation tracking signals and indirect reciprocity are bleak. This shows that the introduction of cheap-talk is detrimental to the evolution of cooperation through indirect reciprocity because of reputation-ruining signals. If "moral signals" are to evolve in this setting, we will need to either examine a more complicated model that allows for a way to police the signals or

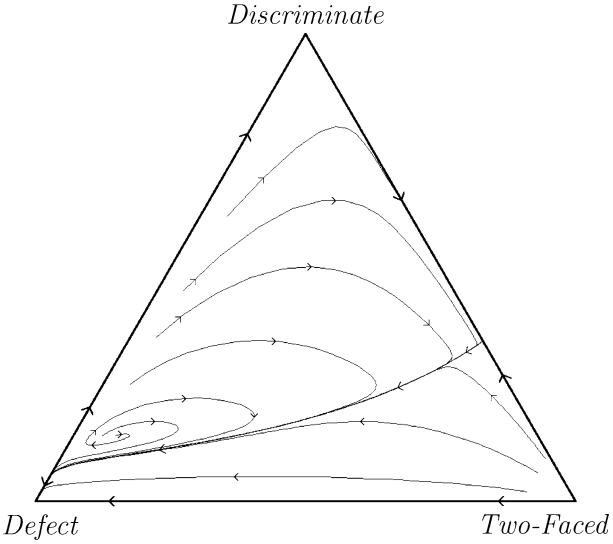

Fig. 4 Problems with the Two-Faced strategy

introduce an additional payoff structure into the signaling portion of the game. These extensions will be explored in the next section.

Including a payoff to signaling

The problems for indirect reciprocity that arise when we explicitly model image tracking by cheap-talk are due to the fact that there is no cost or benefit directly tied to signaling. Individuals are not accountable, in terms of payoffs, for the signals they send regarding others. This is the reason that the discriminator (c,d,0,1) can be invaded by the Two-Faced cooperator (c,c,1,1). We can augment the model above by introducing such a payoff or cost associated with the signals. There are two ways this can be done. First, we can impose a variety of additional payoff (or cost) structures associated with the signaling strategies that are *exogenous* to the setting of indirect reciprocity. Second, we can introduce nuances to the interaction/signaling structure that would allow signaling itself to be rewarded or punished *endogenously* in the setting of indirect reciprocity. We will consider both of these possibilities.

Exogenous payoffs

Here, we consider three possible ways to include an exogenous payoff structure for signals. First, one signal may simply be inherently costly to send. Second, there may be a direct reward associated with using a particular separating strategy such as labeling cooperators with a "0" and defectors with a "1." Third, we can allow for a coevolving, common interest signaling game to serve as the signaling system for representing the images of players. Each of these possibilities will be discussed below.

Suppose that sending signal 1 is costly. There may be many reasons for this: perhaps sending messages about others is difficult or time consuming (and "0" is simply not saying anything), or perhaps there are social consequences for sending signal 1 that we wish to lump into a single exogenous cost for simplicity. Whatever the reason, altering the model in this way will stabilize strategies such as (c,d,0,1) from the Two-Faced cooperator, since (c,c,1,1) requires sending the costly signal.[18] If we simulate evolution in this setting, we see that efficient cooperative states do sometimes arise, but that it is a rare occurrence. For a cost of 0.5 for every use of signal 1, only 6 of 1,000 populations reached cooperative states and all were mixes of (c,c,0,0), (c,c,0,1), (c,d,0,0) and (c,d,0,1). Results in simulations for other cost amounts (ranging from 0.1 to 1.0) are similar. And, for an even higher cost of 2.0, no cooperative outcomes were observed in 500 simulations.

Alternatively, it is possible that there is some external pressure to communicate in a particular way: perhaps as pressure from the linguistic community to conform to an already established signaling system used in other settings. In this case, we can give a one-time benefit x to signaling "correctly" relative to this external standard, which we will stipulate as labeling defectors with a "1" and cooperators with a "0." In simulations, this benefit is given to agents only once each generation and the effect of increasing this benefit is dramatic. With a relatively small one-time benefit of $x = 1$, we find that cooperative populations are stabilized and efficient, but relatively infrequent at approximately 5% compared to the baseline of 79% seen above. Table 2 shows that the proportion of resulting cooperative populations increases as the benefit to correct signaling increases, but even with a substantial bonus of $x = 15$ (the maximum payoff from a given stage game play is 11) it is still much more difficult to get cooperation than in the benchmark case.

The third form of exogenous payoffs that we could introduce involves a coevolving signaling game. The basic motivation behind this is that individuals may be unable to distinguish between image signaling and a different game of common interest signaling as in Fig. 2. Thus, each individual's strategy in the indirect reciprocity setting would double as a strategy for a simple common interest signaling game. Here, there are two states of the world (*left* and *right*) that occur with equal probability. One player observes the state and sends a signal (0 or 1). The other player then observes the signal and acts accordingly (c or d). If c is chosen in the *left* state each get a payoff of y and if d is chosen in the *right* state each get a payoff y, otherwise each receive no payoff.

In this case, we find that there are two cooperative states possible: all (c,d,0,1) or polymorphic mixes involving (c,c,1,0) and (d,c,1,0). The former is optimal, the latter is not. One could interpret these two possible outcomes as different varieties of moral systems. In the first, members assume players are cooperative until defection is detected. In the second, there is a mix of altruists and discriminators who assume others are defectors until they get a signal that says otherwise. Furthermore, they have different signaling systems (or different moral languages)

[18] To avoid negative fitness values, which cannot be computed in the discrete-time replicator dynamics, this "cost" to signal 1 is calculated by giving a small benefit whenever an individual chooses *not* to send the signal. Since the dynamics are governed only by relative fitness, this benefit for some strategies is equivalent to a cost for others.

Table 2 Effects of introducing a bonus to "correct" signaling

Correct signaling bonus	$x = 1$	$x = 5$	$x = 10$	$x = 15$
% of cooperative populations	4.7%	33.7%	43.0%	47.5%

Table 3 Effects of increasing payoff from a coevolving signaling game

Successful signaling payoff	$y = 1$	$y = 5$	$y = 10$
% of cooperative populations	5.8%	31.5%	61.0%

with opposite use of the signals available.[19] Table 3 shows the effect of increasing the signaling payoff y. As before, increasing this payoff increases the proportion of cooperative outcomes.[20]

All three variations of exogenous payoffs for signaling can stabilize cooperation by indirect reciprocity and plausibly allow an image-tracking signaling system to coevolve with cooperation. Interestingly, we have also seen that some exogenous payoffs to signaling may cause different populations to evolve different uses of "moral signals" and corresponding "moral norms." One could, however, question the significance of these results, arguing that introducing an exogenous payoff structure merely pushes the questions of coevolution back: where do such exogenous payoffs come from and why should they apply to cases of indirect reciprocity? This is a fair criticism, and surely whether or not introducing a particular additional exogenous payoff structure is justified will depend on what is being modeled, how we interpret the payoffs, etc. Regardless, one conclusion that can be drawn is that an appropriate exogenous payoff structure for signaling, wherever it comes from, would aid the coevolution of signaling and indirect reciprocity. Perhaps it is important that there be such exogenous pressure for a successful system of "moral signals" to evolve.

To avoid the open questions raised by exogenous payoffs, it is important to also examine models that impose a payoff structure on the signals within the setting of indirect reciprocity itself. The next extension considered will model such an endogenous payoff to the signaling strategies. The idea is that sending signals in a manner different from others may elicit a form of signal-retaliation and the signaler herself may put her image at stake when making claims about others. If this additional complication can allow for stable social signaling which enables cooperation through indirect reciprocity, we will have at least one setting where exogenous payoffs are not required for the evolution of a simple system of "moral signals."

[19] The difference in these two systems is due to the different signaling conventions. The reason the outcomes are asymmetric is due to the asymmetry in the game mentioned above: all players begin with an image of "0."

[20] The situation is somewhat different than before. As the payoff y gets very large, the game of indirect reciprocity will cease to matter and the signaling game will simply dominate determination of fitness.

Endogenous payoffs to signaling

On the standard picture of indirect reciprocity, acting cooperatively will elicit future cooperation from others. It is possible that signaling a certain way may be something that also elicits future cooperation (or defection). To explore this possibility, we will introduce an additional layer of signaling into the model. As before, agents meet in one-shot games and send signals regarding the actions of their opponents. Now, an interaction between two agents (say A and B) as well as the signal sent (by B) are observed by another member of the population (C) with probability q. C may then send a signal regarding B, changing B's image. If B's signal matches what C would have sent regarding A, C labels B as a cooperator ("0" or "1," whatever that would be for C). If B's signal did not match, C gives B the label that corresponds to defection. The idea here is that if agents signal according to your strategy, they are seen as cooperative, and if they signal differently, they are seen as uncooperative. Can this new element of the model stabilize cooperation and the signaling system necessary for image-tracking? Simulations bring both good news and bad news.

The good news is that, with $q = 0.1$, simulations show that the answer here is yes! This new element does create an endogenous benefit which serves to eliminate strategies that created problems for cooperation, such as the Two-Faced strategy. The cooperative populations that arise are a stable mix of discriminators and altruists, and tend to agree on signaling.[21] Moreover, all the resulting cooperative populations are optimal. The only other stable states seen in simulations are mixes consisting of various defecting strategies. This result reveals that it is *possible* for an image-tracking signaling system ("moral signaling") to evolve in the context of indirect reciprocity, for that system to make indirect reciprocity possible, and for that system to be the mechanism for its own enforcement. However, there is reason to be cautious before concluding much more.

The bad news is that, while this is possible, it does not appear likely. The resulting populations reaching cooperative states (with $N = 5$ and $q = 0.1$) was just under 2% of the total simulated populations.[22] The benefit to endogenous signaling is not significant enough to push more than a small number of populations to cooperative states. Thus, this way of making the payoffs to signaling endogenous reveals a *possible* coevolution of signaling and indirect reciprocity. But, without some additional exogenous pressure on signaling, cooperation is far less likely to evolve than in the benchmark image scoring setting.

[21] Some of the members in the stable cooperative populations send the "cooperative" signal of "0" regardless of opponent action. But, in these populations these strategies are never seen deviating from others, since everyone is in fact cooperating.

[22] There are various ways to increase the prospects for cooperation here. For instance, decreasing the benefit to defection to 0.1 (rather than 1) increases the proportion of cooperative populations to 30%. Another possibility is to increase the number of rounds of play. But in any case, there is a large decrease relative to the baseline case involving only (c,c,0,1), (c,d,0,1), and (d,d,0,1). Increasing q does not have a large effect.

Possible extensions

There are several possible variants on the model presented in the previous section that would, intuitively, improve the prospects for the coevolution of "moral signals" and indirect reciprocity. A few of these are briefly described below along with simulation results.

Evolution of behavior in Stag Hunts or Coordination Games in standard settings is much more likely to result in cooperation than in the Prisoner's Dilemma. Using one of these games as our base game in the model above does make the evolution of cooperation more likely. However, the signaling portion of the model has virtually no effect on the evolution; and the resulting populations do not favor one signaling strategy over another. The reason is obvious: in these games, there is no need to "enforce" cooperation once the population is cooperative because it is strictly harmful to behave in another way. Therefore, there is no need to track reputation for enforcement purposes.

Another possibility that could affect the evolution of cooperation and "moral signals" is a bias toward "truth-telling" in the initial populations.[23] Introducing such a bias will, of course, have an effect. However, a small bias only slightly increases the basin of attraction, and a relatively large bias is needed to create substantial changes. Even with truth-telling strategies being, on average, five times more likely than other strategies, only 23.6% of simulations resulted in cooperation. This means that if an initial bias is to help the coevolution of cooperation and "moral signals" it will have to be quite large, which cannot be justified within the scope of this model.[24]

Finally, perhaps there are restricted strategy sets, which can be justified, that will dramatically increase the likelihood of the coevolution of cooperation and "moral signals." Making such modifications certainly has an arbitrary feel, but even if they are justifiable, such restrictions typically do not have a dramatic effect. For example, if we exclude any strategy that responds to "0" with d and "1" with c as well as any strategy that signals "1" for c and "0" for d, we still see only a very small proportion of cooperative signaling populations evolve (only 6 of 500 simulated populations).

In general, the result that the coevolution of indirect reciprocity and moral signaling is unlikely (relative to the benchmark case) seems relatively robust with respect to variations in the model. This can be taken as providing indirect theoretical support, in at least these simple settings, for the following claim: if language is to play the role of reputation tracking, the selective pressure for its evolution must come from outside the arena of indirect reciprocity.

[23] The initial distribution in the strategy space is chosen randomly by selecting a number r_j at random from the interval (0,1) for each strategy j and then setting the proportion of type i in the population to be $x_i = -\ln(r_i)/\sum_j -\ln(r_j)$. An initial bias in the direction of truth-telling can be represented by substituting $-c\ln(r_i)$ for $-\ln(r_i)$ if i is a "truth-telling" strategy where $c > 1$ is a constant representing the strength of the bias. For the results presented here $c = 5$.

[24] Bias may also be introduced into the dynamics itself in the form of a conformist bias (Skyrms 2005). Exploring different dynamics in this setting would be an interesting topic for future work.

Conclusion

The coevolution of signaling and indirect reciprocity is not straightforward. The simple models examined here have provided a starting point for investigating the evolution of "moral signals" and the role they play in our moral systems. The most direct method of modeling such coevolution reveals that some method of policing the use of signals will be needed; without such a method, troublesome strategies such as the "Two-Faced" cooperator may destabilize cooperative populations.

We can introduce an *endogenous* payoff to signaling, which shows that it is possible for an image-tracking signaling system ("moral signals") to coevolve with indirect reciprocity. In this case, however, cooperation is not likely to evolve and it seems improbable that a coevolution of signaling and indirect reciprocity would occur without some additional *exogenous* payoff structure directly related to the signaling. Thus, for the use of "moral signals" to evolve in this setting, it is important that signals be subject to selective pressure apart from indirect reciprocity. Furthermore, when such exogenous pressures are included, interesting possibilities arise. Different populations can evolve different "moral systems" that vary in the way they use "moral signals." These simple models show that even if moral language plays a crucial role in cooperation through indirect reciprocity, it does not follow that this setting has provided the selective pressure driving the evolution of a moral language.

More generally, these results have implications for the claim that *the benefit of large-scale cooperation was a major selective force in the evolution of social communication*. We found that either (i) "moral signals" can evolve but only as an unlikely accident of the starting point of the evolutionary process, or (ii) "moral signals" are likely to evolve but only when there is selective pressure from outside the arena of indirect reciprocity. In either case, these simple models cast doubt on the claim above. If the claim is to be supported theoretically, an alternative model of the evolutionary process will need to be provided. Whether or not such a model can be provided is an open question, but the simple models examined here have highlighted the difficulties involved in giving a positive account.

Acknowledgments I would like to thank Brian Skyrms, Simon Huttegger, Samuel Bowles, Karthik Panchanathan two anonymous referees and the members of the Social Dynamics seminar at UCI for helpful feedback on this paper. Generous financial support was provided by the Institute for Mathematical Behavioral Sciences and the School of Social Sciences at UCI.

References

Aiello LC, Dunbar RIM (1993) Neocortex size, group size, and the evolution of language. Curr Anthropol 34(52):184–193

Alexander RD (1987) The biology of moral systems. Aldine de Gruyter, New York

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

Barrett JA (2008) Dynamic partitioning and the conventionality of kinds. Philos Sci 74:527–546
Brandt H, Sigmund K (2005a) The good, the bad, and the discriminator—errors in direct and indirect reciprocity. IIASA Interim Report IR-05-070
Brandt H, Sigmund K (2005b) Indirect reciprocity, image scoring and moral hazard. Proc Natl Acad Sci USA 120(7):2666–2670
Dunbar RIM (1993) Coevolution of neocortical size, group size and language in humans. Behav Brian Sci 16:681–735
Dunbar RIM (1996) Grooming, gossip, and the evolution of language. Harvard University Press, Cambridge
Fishman MA (2003) Indirect reciprocity among imperfect individuals. J Theor Biol 225:285–292
Harms WF (2000) Adaptation and moral realism. Biol Philos 15:699–712
Harms WF (2004) Information and meaning in evolutionary processes. Cambridge University Press, Cambridge
Harms WF, Skyrms B (2008) Evolution of moral norms. In: Ruse M (ed) Oxford handbook on the philosophy of biology. Oxford University Press, Oxford
Huttegger SM (2007a) Evolution and the explanation of meaning. Philos Sci 74:1–27
Huttegger SM (2007b) Evolutionary explanations of indicatives and imperatives. Erkenntnis 66:409–436
Joyce R (2007) The evolution of morality. MIT Press, Cambridge
Kandori M (1992) Social norms and community enforcement. Rev Econ Stud 59:63–80
Leimar O, Hammerstein P (2001) Evolution of cooperation through indirect reciprocity. Proc R Soc Lond B 268:745–753
Lewis D (1969) Convention. Harvard University Press, Cambridge
Milinski M, Semmann D, Bakker TCM, Krambeck H-J (2001) Cooperation through indirect reciprocity: image scoring or standing strategy? Proc R Soc Lond B 268:2495–2501
Millikan RG (2005) Language: a biological model. Oxford University Press, Oxford
Nakamaru M, Kawata M (2004) Evolution of rumours that discriminate lying defectors. Evolut Ecol Res 6:261–283
Nowak MA, Sigmund K (1998a) The dynamics of indirect reciprocity. J Theor Biol 194:561–574
Nowak MA, Sigmund K (1998b) Evolution of indirect reciprocity by image scoring. Nature 393:573–577
Nowak MA, Sigmund K (2005) Evolution of indirect reciprocity. Nature 437:1291–1298
Ohtsuki H, Iwasa Y (2004) How should we define goodness?—reputation dynamics in indirect reciprocity. J Theor Biol 231:107–120
Panchanathan K, Boyd R (2003) A tale of two defectors: the importance of standing for evolution of indirect reciprocity. J Theor Biol 224:115–126
Skyrms B (1996) Evolution of the social contract. Cambridge University Press, Cambridge
Skyrms B (2002) Signals, evolution and the explanatory power of transient information. Philos Sci 69:407–428
Skyrms B (2004) The stag hunt and the evolution of social structure. Cambridge University Press, Cambridge
Skyrms B (2005) The dynamics of conformist bias. Monist 88:259–269
Sterelny K (2003) Thought in a hostile world. Blackwell, Oxford
Sugden R (1986) The economics of rights, cooperation and welfare. Blackwell, Oxford
Taylor P, Jonker L (1978) Evolutionary stable strategies and game dynamics. Math Biosci 16:76–83
Trivers RL (1971) The evolution of reciprocal altruism. Q Rev Biol 46:35–57
Weibull JW (1995) Evolutionary game theory. MIT Press, Cambridge
Zollman K (2005) Talking to neighbors: the evolution of regional meaning. Philos Sci 72:69–85