For *Oxford Handbook on the Philosophy of Biology* ed. Michael Ruse
draft Oct. 20, 2007

EVOLUTION OF MORAL NORMS

William Harms
Brian Skyrms

Moral norms are the rules of morality, those that people actually follow, and those that we feel people *ought* to follow, even when they don't. Historically, the social sciences have been primarily concerned with describing the many forms that moral norms take in various cultures, with the emerging implication that moral norms are mere arbitrary products of culture. Philosophers, on the other hand, have been more concerned with trying to understand the nature and source of rules that all cultures ought to follow, with relatively little regard for what people actually do. The tension between the two approaches has to do with whether there are any standards higher than the whims of culture in determining right and wrong. Typically, the social sciences say "no", pointing at the diversity of moral beliefs. Most philosophers (along with people of moral conviction) feel that there must be some deeper source of morality than the trends and fads of culture. Unfortunately, the nature and source of such standards has remained something of a mystery. Recent work on the evolution of norms has changed this picture dramatically.

Evolutionary explanation of the emergence of moral norms proceeds in stages, as the evolutionary process itself proceeds in stages, rather than arriving all-at-once at a finished product. First, one must give an account of how behavior in consonance with the norms can arise. This may be no small matter, since some norms prescribe behavior that appears, on the face of it, to reduce Darwinian fitness. The explanations may be different for different classes of norms.

Then, once the evolution of the behavior has been explained, the question remains as to the evolution of its normative status. Part of what makes a kind of behavior a norm consists in society's reaction to those who do not follow the norm. Enforcement marks a norm. Violation of a norm elicits various kinds of enforcement behavior - disapproval, punishment, ostracism - often at some costs to the enforcers. The evolution of these higher-order patterns of behavior must also be explained.

Finally, one might hope in the end to also have an account of the evolution of the language of moral judgment, together with an evolutionary account of its meaning. That is asking for quite a lot, but an evolutionary account of the emergence of moral protolanguage, or of moral signals, would be a beginning.

Any survey of the literature will show that evolutionary game theorists, seeking to explain the evolution of altruism, or cooperation, or fairness, have spent the most time on the first stage of explanation and ethical theorists who are critical of the whole enterprise have spent most of their time on the last stage. Gibbard (1990) is an exception who works to reconcile evolutionary game theory with traditional ethical concerns. Despite some polarization of the field, there is positive constructive work addressing all three stages of the process. In what follows we will attempt to draw a picture of the state of the discussion.

N.B. It should go without saying, that by talking about evolution we include cultural evolution. Many of the tools of evolutionary theory - equilibrium analysis, replicator dynamics, spatially explicit models, meta-population theory - apply equally as well or better to cultural evolution as to genetic evolution. Anyone looking for genetic determinism should look elsewhere.

I. Behavior in Accord with Norms

There are different norms for different kinds of social interaction - norms of justice, norms of cooperation, norms prescribing various kinds of altruistic behavior. The difficulty in understanding their evolution is that the systems in which they evolve are extremely complex, and the forms they take are unique in each instance. Intuitions fail when it comes to predicting the behavior of these sorts of complex dynamical systems. Consequently, in order to make any progress in analyzing these norms, we need simple archetypical models of classes of social interaction to which various norms apply. That is to say that we need the tools of evolutionary game theory.

Game theory is the study of strategic interaction -- situations that involve more than one individual where the action of each affects the other(s). A simple game is specified by the payoffs that result from interaction of possible actions of two or more individuals. What do these payoff numbers mean? In the theory of rational choice, they indicate the perceived benefits to each individual, which determine the rationality of each option. Evolutionary game theory uses games to characterize the forces driving the propagation or dissemination of traits and behaviors. In a context of genetic evolution, payoffs indicate Darwinian Fitness - expected number of progeny. In the context of cultural evolution they should be taken as whatever good that drives differential imitation or other cultural dynamics. For individuals living near the edge of starvation both these may correlate well with food, but this may no longer be true when food is plentiful. In contexts of cultural evolution the interpretation of payoffs is non-trivial, including factors as diverse as the perceived promotion of social status and the satisfaction of acquired tastes.

We will examine three simple games illustrating issues relating to altruism, cooperation, and distributive justice.

Self-sacrificing behavior or **Altruism** is commonly observed in nature, but is a bit of a puzzle since it would seem to reduce Darwinian fitness and thus be eliminated by natural selection. The evolutionary stability of altruism has been extensively discussed in terms of the so-called **Prisoner's Dilemma**. In the simplest case, the Prisoner's Dilemma is a two person "bimatrix" game where each player has two options, Cooperate and Defect. The game, as it is defined, does not require any specific set of payoffs. Instead, a Prisoner's Dilemma occurs anytime certain abstract requirements are met. What is essential is that mutual advantage is maximized if both individuals cooperate, but no matter what others do a given player does better by defecting. This sort of structure can be instantiated by an interaction of any number of players (a social dilemma), but the essential points can be illustrated in a two-player Prisoner's Dilemma an example of which follows. (The pairs of numbers represent the payoffs first for the Row player and second for the Column player.)

	Cooperate	Defect
Cooperate	3,3	1,4
Defect	4,1	2,2

If Row decides which row we play, and Column which column, then we can read the payoff matrix in the following fashion: No matter whether Column cooperates or defects, Row does better by defecting, placing us in the "defect" row. The same reasoning applies to the Column player, so that if both players are rational, we will end up at defect/defect. In such a situation, the act of defecting is said to *strictly dominate* that of cooperating. The only point where no one would do better by changing her action - the only *Nash equilibrium* - is the one at which all players defect. On the other hand, if both were to cooperate both would get a payoff (3) superior to that gained if both defect (2). An *altruistic* player might give up the advantage to be gained by defecting, losing 1, and cooperate to help the other player gain 2. If both players are altruistic they both end up with a payoff of 3, but defecting against an altruist pays better (4).

Social beings can perform many acts which cost them little and benefit others more, like warning of danger, sharing excess food, or merely leaving another's property alone. An act which costs one unit of fitness to perform, and benefits the recipient two, generates a Prisoner's Dilemma. As such, the Prisoner's Dilemma has come to be a theoretical microcosm for the study of the stability of cooperation and its benefits and thus for the evolution of moral behavior. Since cooperation is so commonly observed in nature, the theoretical challenge is to come up with solutions to the Prisoner's Dilemma, plausible modifications of the situation such that cooperation can stabilize. (We will discuss this shortly.)

It should be evident that the names of games do not indicate their subject matter. Rather, they invoke emblematic situations that share the general payoff structure of the game. **Prisoner's Dilemma** takes its name from an imaginary situation in which partners in crime are individually interrogated and offered deals testify against each other, but the game is not about Prisoners but about altruistic cooperation. Similarly, the **Stag Hunt** envisions us faced with the choice of whether to hunt hare or hunt stag, the idea being that hunting hare can be done alone and guarantees modest return, whereas hunting stag offers a bigger return but depends on others showing up to help out – a riskier venture which typifies different sorts of situations than the PD.

The Stag Hunt raises issues of cooperation in a less extreme setting than that of the Prisoner's Dilemma. There are two *Nash Equilibria* in this game, one of which maximizes mutual advantage but which is risky and another which is safe but not so rewarding. The problem is no longer that cooperation is not an equilibrium, but that there is another equilibrium that may be easier to fall into. Here is an example of such a two-player Stag Hunt:

	Cooperate (Stag)	Go it Alone (Hare)
Cooperate (Hunt Stag)	4,4	0,3
Go It Alone (Hunt Hare)	3,0	3,3

The equilibrium where both cooperate gives each the highest possible payoff - it is called the *payoff dominant* equilibrium. But a cooperator faces a payoff of zero if his potential partner decides to chase a Hare instead of helping with the Stag hunt. A hare hunter faces no risk whatsoever, as her payoff does not depend on what the other does. A player who believed that the other was as likely to do one thing as another would opt to hunt hare. In this case the non-cooperative, go-it-alone equilibrium is said to be *risk dominant.* If both players trust one another to cooperate, they will cooperate and achieve the maximum payoff. If neither trusts the other, their lack of trust lands them in the sub-optimal, risk-dominant equilibrium.

"Distributive justice" in philosophical discussion refers not to the justice of basic rights or to the justice of retribution, but to how common goods ought to be distributed among members of a community. Principles of **distributive justice** are illustrated by **bargaining games**. Some commonly held good is to be divided. In the simplest case the payoff is just the amount of good gotten, although in more general cases payoffs to different players can be different functions of the amount of good each receives. Consider the two-player Nash bargaining game called Divide-the-Dollar. There is a commonly held sum of money to be divided among the two players. In all relevant aspects except their identity, the two are indistinguishable.
Their payoff is just the amount that they get.

Each player simultaneously chooses a fraction to demand. If the fractions add up to more than one, there is no bargain and neither gets anything. Otherwise each gets what he demands. There are now an *infinite number of Nash equilibria*. If Row demands a and Column demands b and a + b=1, this is a Nash equilibrium, where no-one can gain by changing strategy. (There is also an odd equilibrium where each demands all and no one gets anything.) All these equilibria (except the odd one just mentioned) use up all the good. So no one could be made better off without someone being made worse off. But almost everyone would agree that in the highly symmetrical conditions stipulated the only fair distribution is the equal split. The principle challenge here also is to introduce plausible conditions where evolution favors the equal split. Such conditions then can be treated as historical hypotheses concerning the origins of our own preference for the equal split. Is this why we cooperate the way we do?

A good theory of evolution of norms might start by explaining the evolution of altruism in Prisoner's Dilemma, of Stag Hunting, and of the equal split in the symmetric bargaining game. *These are not well-explained by classical game theory based on rational choice*. From a technical point of view, they present different theoretical challenges. In the bargaining game, there are an infinite number of equilibria with no principled (rational choice) way to select the cooperative one. In Stag Hunt there are only two, but the non-cooperative one is selected by risk-dominance. In Prisoner's Dilemma the state of mutual cooperation is not a Nash equilibrium at all, and cooperation flies in the face of the rational-choice principle that one does not choose less rather than more. In contrast to rational choice theory, the most common tool of evolutionary game theory is the *replicator dynamics*, in which the propagation rate of each strategy is determined by its current payoffs. These dynamics have a rationale in both biological and cultural evolutionary modeling, and sometimes tell us things that rational choice theory does not.

Evolutionary game theory has, over the last forty years, produced a remarkable literature consisting of contributions from fields as various as philosophy, economics, psychology, ecology and evolutionary biology. By far the greatest effort has been spent on the Prisoner's Dilemma. One of the first ideas introduced was that of *kin selection* [Maynard Smith(1964), Hamilton(1964)]. The idea is that helping one's kin may spread their genes, which may with some likelihood, be one's own genes. Thus, under the right conditions, "selfish" genes may produce altruistic individuals. If treated with care, this idea makes sense and explains a substantial amount of cooperation in nature. Hamilton's (1964) exposition on "inclusive fitness" already makes clear that what is fundamental is the correlation produced by kinship rather than kinship itself. [Hamilton (1964) (1996), Grafen (1984) (1985)]

Robert Axelrod (1984), following game theorists since John Nash, introduced indefinitely *repeated interactions*. Axelrod considers a series of Prisoner's Dilemmas between two individuals. There is a fixed, constant probability of another interaction, no matter how many interactions have already taken place. If this probability is high enough, there may be an equilibrium in which everyone always cooperates. This can happen if both players play the strategy *grim trigger* (Cooperate until the other defects, and then defect forever) or Anatol Rapoport's strategy *tit-for-tat* (Cooperate initially and then do to him what he did to you last time.)

Axelrod and Hamilton (1981) had suggested that altruistic behavior may have gotten started by kin selection, and then spread to non-kin via repeated interactions. A population all playing *tit-for-tat* are at a neutral evolutionary equilibrium. No one playing a different strategy can do better against the natives. That equilibrium is not, however, evolutionarily stable in the sense defined by Maynard Smith which requires that natives do better than any possible mutants. Other strategies (e.g. always cooperate) do equally well. *Tit-for-Tat* resonates with Robert Trivers' 1971 examination of reciprocal altruism in biology, but the strong probabilistic assumptions behind the theory of repeated games leave lingering questions.

It has always been clear that a group of cooperators who interacted primarily with each other would outperform groups of defectors, and so there are various *group selection* models of the evolution of altruism in the Prisoner's Dilemma, where selection is said to operate on groups as well as individuals. Ironically, the first precise group-selection model, the Haystack model, was given by Maynard-Smith (1964) as a criticism of group selection. Maynard-Smith's point was that the model required such special circumstances to produce evolution of altruism that it was unlikely that such a process could be a major explanation of mutual aid in the animal kingdom. When we consider cultural evolution, some group selection models do not seem so implausible. [See Sober and Wilson (1998), Kerr and Godfrey-Smith (2002), Bergstrom (2002), Canals and Vega Redondo (1998), etc] If however, group selection is taken in the most general sense and kin selection is taken in the most general sense, then – as Price (1972) showed, these are mathematically equivalent. [Frank (1995a) (1998), Gardner and West (2004)]

In nature, local interaction is sometimes mingled with kin selection, when individuals live near their kin but, as Hamilton already noted, local interaction in itself can create conditions more favorable for cooperation. In various models where individuals interact with neighbors on a grid, line, circle, or some other spatial structure, and where individuals update their strategies by imitation, it has been shown that cooperation can coexist with defection. [Pollack (1989), Bergstrom and Stark (1993), Grim (1995), Hegselmann (1996), Eshel, Samuelson and Shaked (1998)] Sometimes the dynamics of coexistence can be very complicated. [Nowak and May (1992)] The key to these models is the clustering of cooperators. Those in the interior of a cluster are shielded from defectors and do very well.

The overall implications of spatial structure for cooperation, however, are not entirely unambiguous. Depending on the model, spatial structure may lead to competition with kin within the local community canceling out the effects of cooperation. [Taylor(1992)]

Improvements in computational resources and techniques allow us to model the stability of cooperation in heterogeneous spatially explicit environments, with populations of varying size which exhibit biological rather than cultural propagation patterns. Using such techniques, Harms (2001) showed that a gradient of environmental hostility alone can segregate cooperators and defectors sufficiently to stabilize cooperative behavior in one-shot Prisoner's Dilemma. Similar techniques can be used to model the interaction between biological and cultural evolution.

All of the foregoing secure the possibility of cooperation in Prisoner's Dilemma by generating, in one way or another, a positive correlation of types - cooperators and defectors meet their own type more often than would be expected with random encounters. [Hamilton (1964) (1996), Eshel and Cavalli-Sforza(1982), Sober (1992), Skyrms(1996)] Positive correlation helps with Stag Hunt and Nash Bargaining, too. If players in a Stag Hunt meet their own type, cooperators (Stag Hunters) do better than loners (Hare hunters). If bargainers meet their own type, those demanding an equal split do better than others. For instance, Alexander (2000) investigates a local interaction model of bargaining, and shows how it leads to the equal split. Zollman (2005a) shows how a local interaction model combined with signaling leads to cooperation in the Stag Hunt game. Some additional mechanisms, however, are of interest in these games.

Positive correlation between types is not the whole story, however. In the bargaining literature, we find that the concept of a *stochastically stable equilibrium* [Foster and Young (1990), Kandori, Malaith and Rob (1993)] uniquely selects the Nash bargaining solution, which in our Divide-the-Dollar game is the equal split [Young (1993a). (1993b)]. In these models, evolution takes place according to the replicator dynamics, and there is no positive correlation mechanism insuring that players are more likely to play with their own kind. Instead, the population is perturbed by random shocks, or "mutations" in a finite population version, with very small probability. In the long run, the population spends most of its time in a state where all players demand half. This does not help in Stag Hunt, where the stochastically stable equilibrium is the Hare-hunting one (nor in Prisoner's Dilemma, where the only equilibrium is mutual defection.)

Finally, we are just beginning to fully understand the role learning may play in stabilizing cooperation. For instance, in the Stag Hunt, if Stag Hunters can learn to interact with each other, Stag Hunting can flourish. If social networks are dynamic, and evolve by reinforcement learning, this effect can favor Stag hunting. [Skyrms and Pemantle (2000)] Much the same effect is achieved if players are able to move to "islands" where Stag Hunters are isolated. [Ely (2002), Oechssler (1999), Diekemann (1999)]. This may not be so easy in Prisoner's Dilemma, because Defectors want to interact with Cooperators just as much as Cooperators do. In the pure Stag Hunt game, Hare hunters do not care. Correlation induced by learning is much more complicated in bargaining games, and all sorts of things can happen [Skyrms (2004)].

II. Enforcement and Punishment

The preoccupation of theorists with the first stage of the evolution of norms is understandable. New mechanisms for the stability of cooperation continue to be discovered, and the clarity of the basic games has facilitated productive research across traditional disciplinary boundaries. Understanding the evolution of enforcement behavior provides a more difficult challenge, and one in which theoretical expectations are commonly confounded by empirical evidence. We will concentrate on our own species, but note that enforcement and punishment are not unknown in other species. [Clutton-Brock and Parker (1995), Ratnieks and Visscher (1989)] Punishment is more common than reward in nature, perhaps because the costs of punishment decline as the equilibrium it supports is approached, while the net costs of rewarding equilibrium behavior increase.[Gardner and West (2004)]

In the Ultimatum Game [Güth, Schmittberger and Schwarze (1982)], players divide a windfall with the twist that Player 1 offers a share to Player 2, keeping the rest. Player 2 accepts or rejects this share and if rejected neither player gets anything. Thus player 2 can punish player 1 at a cost to himself. As a simple problem of rational choice, Player 2 should accept any non-zero offer (since something is better than nothing), and Player 1, knowing this, should therefore make the smallest non-zero offer possible, in order to maximize her own payoff. Presumably, Player 2 would reason similarly if the situation were reversed. Of course, this is not how humans actually behave.

Experiments done on college students at universities around the world have shown a strong bias toward equal split offers or offers reasonably close to these. If player 1 asked for a very large proportion of the good, player 2 would often reject the offer leaving both with nothing. It would appear that real life Player 2's feel the need to punish Player 1's they perceive as being greedy or taking advantage of their position of power. The fact that player 2's who punish experience anger at being made what they regard as a derisory offer is disclosed by post-play questions and has even been confirmed by brain imaging. Player 1's who anticipate this make offers which are closer to 50%. At this point, rational choice theory can say little other than that people have a strong preference for fairness, but the source and justification of that preference remains a mystery. Hopefully evolutionary analysis can do better.

The first thing to ask is whether the observed behavior can evolve in the simplest evolutionary model, replicator dynamics with random encounters and a small amount of mutation. (Recall that in the replicator dynamics, a strategy increases or decreases according to how well it does at each moment in the game.) For a two population model - one population of player 1's and one population of player 2's - an affirmative answer is given by Binmore, Gale and Samuelson (1995). In a one population model, in which players are sometimes in the role of player 1 and sometimes in the role of player 2, an affirmative answer is given in Harms (1997). These involve polymorphisms stabilized by a balance of small selection pressures in one direction and small net mutation forces in the other. A rational choice theorist could say that utility functions have evolved, and a number of theorists have explored this point of view. The idea that emotions have evolved to guarantee threats (and promises) was put forward by Hirshleifer (1987) and Frank (1988).

The positive results of Binmore, Gale and Samuelson, and of Harms are that the behavior we have described *could possibly* evolve, not that it must. In fact, a number of results are possible depending on details of initial conditions and the structure of mutation probabilities. This suggests that we might also find other equilibria instantiated in different societies. In fact, in studies of twenty small scale societies anthropologists have found all kinds of ultimatum game behavior. [Henrich, Boyd, Bowles, Camerer, Fehr, and Gintis (2004)] They find that other aspects of the societies influence the types of norms observed in ultimatum games. This means that the correct evolutionary analysis must be more complicated, with a variety of interactions driving the evolution of norms general enough to be applied across the board. Zollman (2005b) analyzed evolution in a mixed series of Nash bargaining and ultimatum bargaining interactions, and found that evolution of the equal split is more likely in the mixed environment that in either pure environment.

In a somewhat richer context, experimentalists have found people willing to punish in *public goods* provision games. The basic game (without punishment) proceeds as follows. All individuals are given an initial monetary endowment. They can either contribute none, some, or all of it toward a public good – something like streets or law enforcement to which everyone contributes and which benefit everyone. What they don't contribute they get to keep. The money in the public pot is multiplied by some constant greater than 1, and the result is distributed equally among members of the group. If everyone contributes everything then they get back what they contributed multiplied by the constant. However, if someone doesn't contribute he gets to keep all his money and gets a share of the amplified contributions of everyone else. The multiplier and the number in the group are arranged so that contributing nothing is the strictly dominant act for each player. (No matter what others do, a player's personal payoff is higher if he contributes nothing.) Experiments with this basic game show that individuals enter with many making substantial contributions, but these numbers fall with repeated play. [Ledyard (1995)].

The basic game is then modified so that after the contributions to the public pot are made, subjects find out what others contributed and have an opportunity to punish [by fining] others at some cost to themselves. The basic experimental results are (i) that subjects are willing to fine free riders (ii) that contributions to the public good are high and, if anything, get higher as the game is repeated. [Ostrom, Walker and Gardner (1992); Fehr and Gachter (2000) (2002)]. Costly punishment is undertaken, and it is effective. The theoretical treatment of public goods provision with punishment in the basic replicator dynamics model will go much like that of ultimatum bargaining. And like the case of ultimatum bargaining, there is a strong likelihood that norms applied here evolved in a more general setting. Brandt, Hauert and Sigmund (2003) give a local interaction model with three-player interactions in which punishment can evolve and enforce cooperation. This effect is amplified if reputation effects are added. For a game close to the public goods provision game discussed here, Boyd et al. (2003) give a group-selection model in which punishment evolves. See also Fowler (2005) for another model of evolution of punishment in public goods provision games, and Hauert, De Monte, Hofbauer and Sigmund (2002) for the possibility of evolutionary cycles in this context.

We should also remind ourselves that we have already seen a form of punishment already in section I. In repeated Prisoner's Dilemma, Grim Trigger is a punishing strategy. Once someone defects, Grim Trigger never cooperates with him again. This punishment strategy may not carry an explicit cost, but there is an implicit opportunity cost if there is any likelihood that potential cooperation is thereby foregone. Tit-for-Tat offers much more mild punishment, and there are all kinds of punishing strategies - deterministic and probabilistic - in between. This kind of punishment forms the basis of “folk theorems” of repeated game theory, which show in general how punishment can sustain socially efficient equilibria. [Fudenberg and Maskin (1986)] The original folk theorems where proved in the context of repeated interactions between two players. They were extended to community setting by Sugden (1986), Kandori. (1992) and Nowak and Sigmund (1998). For a comparison of different ways of dealing with reputation in community enforcement, see Ohtsuki and Iwasa (2006). The “opting out” model of Kitcher (1993), the social network dynamics of Skyrms and Pemantle (2000) and that of Santos, Pecheco and Lenaerts (2006), and the indirect forms of partner choice in Ely (2002), Oechssler (1999), Diekemann (1999) all can be thought of as incorporating some form of punishment by ostracism.

Theoretical models of the evolution of enforcement and punishment do not yet exhibit quite the scope and variety that we find in models of the evolution of altruism, cooperation and distributive justice, but the field is very active. Punishment that is costly and that enforces norms of cooperation is sometimes called *altruistic punishment* in this growing literature. This is something of a euphemism, since both the punisher and the individual pay a price, and a more candid evolutionary description might be “righteous spite”. [See Johnstone and Bshary(2004)] It should also be noted that punishment need not entail a net cost to the punisher. If non-cooperators manage to visit free-riders more than punishers in the population, punishment may carry a net benefit. [Gardner and West (2004)] This is a more subtle instance of the importance of correlation for the evolution of cooperation. There are other aspects of punishment still to be explored. Nevertheless, in what has been done, we already see a range of theoretical models of the evolution of effective enforcement of norms by punishment.

III. Moral Signals

Naturalistic accounts of morality are often criticized as unable to account for the normative meaning of moral language and the normative force of moral judgement. The most satisfying completion of an evolutionary account of moral norms would include an account of the meaning of moral statements that speaks to this concern. The evolution of language is itself a very large topic which includes many unanswered questions. But considerable progress has been made in understanding the evolution of signals, and this is surely the place to begin.

The classic signaling game derives from Lewis (1969). Two players, a Sender and a Receiver, must conspire to coordinate an act with a state of the world. Only the Receiver acts, and only the Sender perceives the state. For each state of the world there is a unique mutually beneficial act. Both players get a payoff if the Receiver acts "correctly" and there is no payoff otherwise. Notice that this is a pure coordination game, where the players succeed or fail together. Sender and Receiver are not assumed to have any pregame understanding as to which signal indicates which state. The "meaning" of the signals, if they have one, is created by the players being at an equilibrium of the game.

The simplest signaling games involve two states, two acts, and two signals. Suppose Act1 pays off in State1 and Act2 pays off in State2, and let's call the signals Red and Blue just so there is no suggestive association between signals and states (or acts). This yields four Sender strategies and four Receiver strategies.

Sender Strategies	***Receiver Strategies***
S1) State1: Send Red; State2: Send Blue	R1) Act1 if Red; Act2 if Blue
S2) State1: Send Blue; State2: Send Red	R2) Act1 if Blue; Act2 if Red
S3) Always Send Red	R3) Always Act1
S4) Always Send Blue	R4) Always Act2

S1 and R1, in combination, always convey the information, do the right thing, and get the payoff. S2 together with R2 is an equally effective pairing. These are alternative signaling systems, in which red and blue have different meanings.

In one-population evolutionary games, players take both sender and receiver roles, and so their strategies are conditional: if sender, do this and if receiver, do that. There are 16 such complete strategies: S1R1, S1R2, etc. A population consisting of S1R1 types who used strategy S1 if sender and R1 if receiver would always get things right. Mutants who did something different would get a lower payoff than the natives. That is to say that S1R1 is an *evolutionarily stable strategy.* The same can be said of the alternative signaling system strategy S2R2. In fact, these signaling system strategies are the unique evolutionarily stable strategies in this signaling game. In the replicator dynamics, almost every possible state of the population is carried to one signaling system or another. [For further discussion see Skyrms (1996) (1999) and Huttegger (2005)]

For the purposes of understanding morality, the equilibrium properties of these games are less important than the precise nature of the meaning that attaches to signals at equilibrium. If an evolving system arrives at a state where all players play S1R1, then the established convention is to send Red when State1, and to perform Act1 in response to Red. According to the convention, Red *means* "The world is in State1" but also *means* "Do Act1". It indicates State1 and prescribes Act1.

Signals in the game share with moral utterances ("stealing is wrong") precisely the features that have made the latter theoretically intractable: they possess objective truth conditions (according to the convention) and yet they command action directly, without mediation by consideration of the receivers' needs or desires. It would seem that, from the standpoint of the structure of meaning, our moral intuitions seem more like animal warning cries than either statements or commands.

In nature, species-specific animal warning cries share this kind of primitive semantic content. The alarm calls of vervet monkeys can be viewed as indicating the nature of the predator or as prescribing the correct evasive behavior. This sort of signaling system is found in prairie dogs, meerkats, jungle fowl and domestic chickens. [Cheney and Seyfarth (1990) Hauser(1996), Maynard Smith and Harper 2003)] Millikan (1996) calls such signals “pushme-pullyus”, facing in both indicative and imperative directions. For instance, the beavers’ tail slaps mean at once *that* there is danger and *to* swim under the dam. The meaning of emotions (chemical signals in the brain) can be understood in the same way. Harms (2004) sees this duality as fundamental to the meaning of signals. Instead of taking propositions, in the philosopher's sense, as the theoretical basis of meaning he emphasizes the *primitive content* of signals which consists of both an indicative and imperative aspect. This is where protolanguage begins; the two aspects of primitive meaning are only distinguished somewhere down the line in the development of language.

Warning cries, however, are not normative in the sense at issue (the Beaver isn’t wrong not to go under the dam) so primitive content can not be the complete story of norms. Harms has proposed that the normativity which concerns ethicists emerges when cooperative enforcement controls acquire primitive content. If a dedicated “cheater detection” mechanism (Trivers 1971) were to arise in order to enforce some cooperative convention, then the signal mediating between perception of cheating and enforcement behavior would have primitive content. It could be a signal to the rest of the community, mobilizing community enforcement. [Signals would be a natural addition to the "image scoring" model of Nowak and Sigmund (1998)].

Such an enforcement signal could eventually become internalized. The internalized signal would be "true" when the convention it enforced was violated, so that it would be *about* the historically established convention and an individual’s relationship to it. It would also command directly, just as normative intuitions seem to. If something like this is the case, then the normative force that has so puzzled philosophers is just what it feels like to “believe” a signal with primitive content, and our moral nature may turn out to be something older and more basic than we have imagined. Clearly this is only a beginning of an account of the evolution of moral language and judgment, whose ultimate success depends on future research and its critical evaluation, but together with work on the evolution of cooperation and enforcement points to the real possibility of a materialist theory of norms which avoids the pitfalls of relativism.

References

Alexander, J. M. 2000. “Evolutionary Explanations of Distributive Justice.” *Philosophy of Science* 67: 490--516.

Alexander, J. M. and B. Skyrms 1990. “Bargaining with Neighbors: Is Justice Contagious?” *Journal of Philosophy* 96: 588-598.

Alexander, R. D. 1987. *The Biology of Moral Systems*. New York: de Gruyter.

Axelrod, R. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Axelrod, R. and W. D. Hamilton, 1981. “The Evolution of Cooperation” *Science* 211:1390-1396.

Bergstrom, T. 2002. “Evolution of Social Behavior: Individual and Group Selection Models” *Journal of Economic Perspectives* 16:231-238.

Bergstrom, T. and O. Stark 1993. “How Altruism can Prevail in an Evolutionary Environment” *American Economic Review* 85: 149-155.

Bicchieri, C., J. Duffy and G. Tolle 2004. "Trust Among Strangers" *Philosophy of Science* 71: 286-319.

Binmore, K. 1993. *Playing Fair: Game Theory and the Social Contract I.* Cambridge, Mass.: MIT Press.

Binmore, K. 1998. *Just Playing: Game Theory and the Social Contract II.* Cambridge Mass.: MIT Press.

Binmore, K., J. Gale, and L. Samuelson 1995. “Learning to be Imperfect: The Ultimatum Game” *Games and Economic Behavior* 8, 56-90.

Boyd, R. and Richerson, P. 1985. *Culture and the Evolutionary Process*. Chigaco: University of Chicago Press.

Boyd, R., H. Gintis, S. Bowles, and P. Richerson 2003. "The Evolution of Altruistic Punishment," *Proceedings of the National Academy of Sciences of the USA*. 100, 3531-3535.

Brandt, H., C. Hauert, and K. Sigmund 2003. "Punishment and Reputation in Spatial Pubic Goods Games. *Proceedings of the Royal Society of London Series B.* 270: 1099-1104.

Canals, J. and F. Vega-Redondo 1998. “Multi-level Evolution in Population Games” *International Journal of Game Theory* 27: 21-35.

Carpenter, J. P. 2005. Punishing Free Riders: How Group Size Affects Mutual Monitoring and the Provision of Public Goods. *Games and Economic Behavior* forthcoming.

Cheney, D. L. and R. M. Seyfarth 1990. *How Monkeys See the World.* Chicago: University of Chicago Press.

Clutton-Brock, T. H. and Parker, G. A. 1995. “Punishment in Animal Societies” Nature 373: 209-216.

Danielson, P. 2002. “Competition Among Cooperators: Altruism and Reciprocity” *Proceedings of the National Academy of Sciences of the USA* 99 (Suppl. 3):7237-7242.

Diekemann, T. 1999. “The Evolution of Conventions with Mobile Players” *Journal of Economic Behavior and Organization* 38: 93-111.

Ellison, G. “Learning, Local Interaction and Coordination” *Econometrica* 61: 1047-1071.

Ely, J. 2002. “Local Conventions” *Advances in Theoretical Economics* v.2, n1 http://www.bepress.com/

Eshel, I. and L. L. Cavalli-Sforza 1982. “Assortment of Encounters and the Evolution of Cooperativeness” *Proceedings of the National Academy of Sciences of the USA*: 79: 331-335.

Eshel, I., L. Samuelson, and A. Shaked 1996. “Altruists, Egoists and Hooligans in a Local Interaction Model” *American Economic Review* 88:157-179.

Fehr, E. and S. Gachter. 2000. “Cooperation in Public Goods Experiments” *American Economic Review* 90: 980-994.

Fehr, E. and S. Gachter. 2002. “Altruistic Punishment in Humans” *Nature* 415: 137-140.

Foster, D. and H. P. Young 1990. “Stochastic Evolutionary Game Dynamics” *Theoretical Population Biology* 28: 219-232.

Fowler, J. 2005. "Altruistic Punishment and the Origin of Cooperation" *Proceedings of the National Academy of Sciences of the USA.* 102: 7047-7049.

Frank, R. 1988. *Passions Within Reason*. New York: Norton.

Frank, S. A. 1995a. “George Price’s Contributions to Evolutionary Genetics” *Journal of Theoretical Biology* 175:373-388.

Frank, S. A. 1995b. Mutual Policing and Repression of Competition in the Evolution of Cooperative Groups” *Nature* 377: 520-522.

Frank, S. A. 1998. *Foundations of Social Evolution*. Princeton, N. J.: Princeton University Press.

Fudenberg, D. and E. Maskin 1986. “The Folk Theorem in Repeated Games with Discounting and with Incomplete information.” *Econometrica* 54:533-54.

Fudenberg, D., D. Levine and E. Maskin 1994. “The Folk Theorem with Imperfect Public Information” *Econometrica* 54:533-554.

Gardner, A. and S. A. West 2004. “Cooperation and Punishment, Especially in Humans” *The American Naturalist* 164, 6.

Gibbard, A. 1990. *Wise Choices, Apt Feelings* Oxford: Clarendon Press.

Grafen, A. 1984. "Natural Selection, Kin Selection and Group Selection" In *Behavioral Ecology: An Evolutionary Approach* J. R. Krebs and N. B. Davies eds. 62-84 Sunderland, Mass.: Sinauer.

Grafen, A. 1985. "A Geometric View of Relatedness" In *Oxford Surveys in Evolutionary Biology* v 2 R. Dawkins and M. Ridley eds. 28-89. Oxford: Oxford University Press.

Grim, P. 1995. "The Greater Generosity of the Spatialized Prisoner's Dilemma" *Journal of Theoretical Biology* 172: 363-359.

Güth, W., R. Schmittberger and B. Schwartze 1982. “An Experimental Analysis of Ultimatum Bargaining” *Journal of Economic Behavior and Organization* 3: 367-388.

Hamilton, W. D. 1964. “Genetical Evolution of Social Behavior I and II” *Journal of Theoretical Biology* 7: 1-52.

Hamilton, W. D. 1996. *Narrow Roads of Geneland* San Francisco: W. H. Freeman.

Harms, W. 1997. “Evolution and Ultimatum Bargaining” *Theory and Decision* 42: 147-175.

Harms, W. 2001. “Cooperative Boundary Populations: The Evolution of Cooperation on Mortality Risk Gradients” *Journal of Theoretical Biology* 213: 299-313.

Harms, W. 2004. *Information and Meaning in Evolutionary Processes*. New York: Cambridge University Press.

Hauert, C. S. De Monte, J. Hofbauer, and K. Sigmund 2002. "Volunteering as a Red Queen Mechanism for Cooperation in Public Goods Games" *Science*. 296: 1129-1132.

Hauser, M. D. 1996. *The Evolution of Communication*. Cambridge, Mass.: MIT Press.

Hegselmann,R. 1996. Social dilemmas in Lineland and Flatland. In Liebrand, WBG and D. Messick eds. *Frontiers of Social Dilemmas Research.* Berlin: Springer Verlag 337-362.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E.Fehr, H. Gintis 2004. *Foundations of Human Sociality:Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies.* New York: Oxford University Press.

Hirshleifer, J. 1987. "On the Emotions as Guarantors of Threats and Promises" In *Latest on the Best: Essays on Evolution and Optimality*. Ed. Dupré, J. Cambridge, Mass: MIT Press.

Hirshleifer, D. and E. Rasmussen 1989 "Cooperation in a Repeated Prisoner's Dilemma with Ostracism" *Journal of Economic Behavior and Organization* 12: 87-106.

Huttegger, S. 2005. "Evolution and the Explanation of Meaning" working paper. University of Salzburg. (forthcoming in Philosophy of Science).

Johnstone, R. A., and Bshary, R. 2004. "Evolution of Spite through Indirect Reciprocity" *Proceedings of the Royal Society of London B* 271: 1917-1922.

Kandori, M., G. J. Mailath and R. Rob 1993 "Learning, Mutation and Long-Run Equilibria in Games." *Econometrica* 61: 29-56.

Kerr, B. and P. Godfrey-Smith 2002 "Individualist and Multi-level Perspectives on Selection in Structured Populations" *Biology and Philosophy* 17:477-517.

Kitcher, P. 1993 "The Evolution of Human Altruism" *Journal of Philosophy* 10: 497—516.

Kandori, M. 1992 "Social Norms and Community Enforcement" *Review of Economic Studies* 59:63-80.

Ledyard, J. 1995. "Public Goods: A Survey of Experimental Research" In Kagel, J. and Roth, A. eds. *Handbook of Experimental Economics* Princeton: Princeton University Press, 111-194.

Lewis, D. 1969. *Convention* Cambridge, Mass.: Harvard University Press.

Maynard-Smith, J. 1964. "Group Selection and Kin Selection" *Nature* 201:1145-1146.

Maynard-Smith, J. 1976. "Group Selection" *Quarterly Review of Biology* 51: 277-283.

Maynard-Smith, J. and D. Harper 2003. *Animal Signals.* Oxford: Oxford University Press.

Millikan, R. G. 1984. *Language, Thought, and Other Biological Categories: New Foundations for Realism. Cambridge, Mass.: MIT Press.*

Millikan, R. G. 1993. *White Queen Psychology and Other Essays for Alice* Cambridge, Mass.: MIT Press.

Millikan, R. G. 1996. "Pushmi-Pullyu Representations" in *Mind and Morals: Essays in Cognitive Science and Ethics* Ed. L. May, M. Friedman and A. Clark 145-161. Cambridge, Mass.: MIT Press.

Nowak, M. A. and R. M. May 1992. "Evolutionary Games and Spatial Chaos" *Nature* 359, 826-829.

Nowak, M. A. and Sigmund, K. 1998. "Evolution of Indirect Reciprocity by Image Scoring" *Nature* 393: 573-577.

Oechssler, J. 1997. "Decentralization and the Coordination Problem" *Journal of Economic Behavior and Organization* 32, 119-135.

Ohtsuki, H. and Iwasa, I. 2006. "The Leading Eight: Social Norms that can Maintain Cooperation by Indirect Reciprocity" *Journal of Theoretical Biology* 239:435-444.

Ostrom, E. J. Walker and R. Gardner. 1992. 'Covenants with and without a sword: Self-governance is possible" *American Political Science Review* 86: 404-417.

Price, G. R. 1972. "Extension of covariance selection mathematics." *Annals of Human Genetics* 35:485-90.

Pollock, G.B. 1989. "Evolutionary Stability in a Viscous Lattice" *Social Networks* 11: 175-212.

Rankin, F. W., J. B. Van Huyck and R. Battallio 2000. "Strategic Similarity and Emergent Conventions: Evidence From Similar Stag Hunt Games" *Games and Economic Behavior* 8: 164-212.

Ratnieks, F. L. W. and P. K. Visscher 1989 "Worker Policing in the Honeybee" *Nature* 342: 796-797.

Robson, A. J. 1990. "Efficiency in Evolutionary Games: Darwin, Nash and the Secret Handshake." *Journal of Theoretical Biology* 144: 379-396.

Santos, F. C., Pacheco, J.M. and Lenaerts, T. (2006) "Cooperation Prevails when Individuals Adjust their Social Ties" *PLoS Computational Biology* 2(10): e140.1-6.

Skyrms, B. 1996. *Evolution of the Social Contract* New York: Cambridge University Press.

Skyrms, B. 1999. "Stability and Explanatory Significance of Some Simple Evolutionary Models" *Philosophy of Science* 67: 94-113.

Skyrms, B. 2004. *The Stag Hunt and the Evolution of Social Structure* New York: Cambridge University Press.

Skyrms, B. and R. Pemantle 2000. "A Dynamic Model of Social Network Formation" *Proceedings of the National Academy of Sciences of the USA* 97: 9340-9346.

Sober, E. 1992. "The Evolution of Altruism: Correlation, Cost and Benefit" *Biology and Philosophy* 7: 177-197.

Sober, E. and Wilson, D. S. 1998: *Unto Others: The Evolution and Psychology of Unselfish Behavior.* Cambridge, Mass.: Harvard University Press.

Sugden, R. 1986. *The Economics of Rights, Co-operation and Welfare*. N.Y. Basil Blackwell.

Taylor, P. D. 1992. "Altruism in Viscous Populations: an inclusive fitness model" *Evolutionary Ecology* 6: 352-356.

Trivers, R.1971. "The Evolution of Reciprocal Altruism" *Quarterly Review of Biology* 46:35-57.

Vanderschraaf, P. and J. M. Alexander 2005 "Follow the Leader: Local Interaction with Influence Neighborhoods" *Philosophy of Science* 72: 86-113.

Young, H. P. 1993a "The Evolution of Conventions" *Econometrica* 61: 57-84.

Young, H. P. 1993b "An Evolutionary Model of Bargaining" *Journal of Economic Theory* 59: 145-168.

Young, H. P. 1998. *Individual Strategy and Social Structure.* Princeton, N. J.: Princeton University Press.

Zollman, K. 2005a. "Talking to Neighbors: The Evolution of Regional Meaning" *Philosophy of Science* 72:69-85.

Zollman, K. 2005b. "Evolutionary Explanations in a Complex Environment" working paper UC Irvine.