The Evolution of Spite, Recognition, and Morality

Patrick Forber and Rory Smead*†

Recognition of and responsiveness to the behavioral dispositions of others are key features of moral systems for facilitating social cooperation and the mediation of punishment. Here we investigate the coevolutionary possibilities of recognition and conditional social behavior with respect to both altruism and spite. Using two evolutionary models, we find that recognition abilities can support both spite and altruism but that some can only coevolve with spite. These results show that it is essential to consider harmful social behaviors as both a product of and an influence on the core features of our moral systems.

1. Introduction. A behavior is altruistic when the actor incurs a cost to confer a benefit on another. The self-sacrificial nature of altruism has placed it at the forefront of discussions surrounding the evolution of morality (Alexander 1987; Sober and Wilson 1998; Sterelny 2012; Tomasello and Vaish 2013). Altruism has also captured the attention of theoretical biologists because of the evolutionary puzzle it presents. How could such behavior evolve when it is clearly better, in terms of individual Darwinian fitness, to accept benefits from others but avoid paying the associated costs? Evolutionary solutions to the puzzle of altruism involve, in some way or other, positive assortment among behaviors (Hamilton 1975; Skyrms 1996; Fletcher and Doebeli 2009). If the benefits of altruism tend to flow toward altruists, the self-sacrificial behavior can generate a relative advantage. If we can fully un-

*To contact the authors, please write to: Patrick Forber, Department of Philosophy, Tufts University, Miner Hall, 14 Upper Campus Rd., Medford, MA 02155, e-mail: patrick.forber@tufts.edu. Rory Smead, Department of Philosophy and Religion, Northeastern University, Holmes Hall, 360 Huntington Ave, Boston, MA 02115; e-mail: r.smead@neu.edu.

†Both authors contributed equally to the article. We would like to thank Kevin Zollman for comments on an earlier version of the project as well as the audiences at the PSA 2014 Biennial Meeting and the Carnegie Mellon University Philosophy colloquium series for valuable feedback.

Philosophy of Science, 83 (December 2016) pp. 884–896. 0031-8248/2016/8305-0021$10.00

derstand how nature has solved this puzzle, the thought goes, we will have the beginnings of an account of the evolution of moral systems: social cooperation lays the foundation for developing and enforcing norms of behavior. As Alexander (1987, 1) puts it: "moral systems are societies with rules." To follow and enforce those rules, we must be able to recognize individuals and their behavioral tendencies to act accordingly. One way this can happen is for individuals to behave conditionally: only help those who help or harm those who do not help.[1] In both cases, it is essential that individuals be able to recognize the behavior of other individuals and adjust their own behavior accordingly.

This is only the beginning of an answer to the evolutionary puzzle, however. What different mechanisms can generate the necessary recognition and conditional behavior? Can these coevolve with altruism? How might these mechanisms influence other social behaviors? To investigate these questions, we model two prominent mechanisms for conditional social behavior: exogenous signals deployed before one-shot interactions and reciprocity in repeated interactions. Using these models, we investigate the relationship between these mechanisms and spite, costly behavior that harms others. Our findings are twofold. First, both sets of mechanisms support spite as well as altruism. Second, in some central cases, spite but not altruism can coevolve with recognition and conditional behavior. Insofar as recognition and conditional behavior are central to our accounts of the evolution of morality, it is essential to consider harmful social behaviors as both a product of and an influence on the core features of our moral systems.

2. Altruism, Spite, and Hamilton's Rule(s). We can characterize both altruism and spite in the context of a simple interaction. Suppose there are two individuals and one has an opportunity provide a benefit (b) to the other individual at a cost of (c) to herself. Doing so would be altruistic. Alternatively, individuals may have the opportunity to harm (h) other individuals at a cost (c) to themselves. Given that the benefit (or harm) does not directly affect the actor, both behaviors present evolutionary puzzles (Lehmann, Bargum, and Reuter 2006; West and Gardner 2010): why do such costly behaviors occur in nature?

Genetic relatedness was one of the earliest answers considered. From the gene's eye view, helping family reproduce may be more advantageous than producing offspring. The generalization of this trade-off is known as Ham-

1. While our focus in this article is on conditional behavior, there are many other potential solutions that do not involve conditional behavior, including kin selection (Hamilton 1964), group selection (Sober and Wilson 1998), and spatial interactions (Pollack 1989; Nowak and May 1992).

ilton's rule. As Hamilton came to realize, relatedness need not be interpreted in genetic terms (Hamilton 1964; Price 1970; Skyrms 1996). If the relatedness coefficient is interpreted as positive or negative assortment, then Hamilton's rule can be generalized for both altruism and spite (Smead and Forber 2013). Suppose r^+ represents the degree of positive assortment (increasing the chances of encountering the same behavioral type). Then Hamilton's rule can be generalized to $r^+ > c/b$. There is an analogous rule regarding the evolution of spite. If r^- represents the degree of negative assortment (increasing the chances of encountering a different behavioral type), then we can express Hamilton's condition for spite as $r^- > c/h$.

Conditional social behavior is one common way to facilitate the assortment needed to stabilize altruism: simply make the altruistic sacrifice conditional on the recipient being an altruist, and only altruists receive the benefit. This creates positive assortment among the behaviors even if individual encounters are random (Michod and Sanderson 1985). Of course, how exactly this conditionality works is paramount. We consider two types of conditional social behavior: social behavior that is conditional on some exogenous signal about the type of individual (sec. 3) and social behavior that is conditional on past behavior (sec. 4). In each case there are parallels and differences between the evolution of spite and altruism. And, as we will see, the differences identify a surprising role for spite in the evolution of recognition.

3. Signals and Conditional Social Behavior. One common mechanism of assortment is conditional behavior combined with a phenotypic marker that reliably identifies an individual's type, commonly referred to as a "greenbeard" (Dawkins 1976; Gardner and West 2009). The idea is that altruism could evolve if altruists were able to conditionally direct their helping behavior only toward one another by way of some identifying marker. To illustrate, suppose there are two types in the population: altruistic greenbeards (A) and egoists (E). Type A has an identifiable marker and only provides help (b at cost c) to those with the same marker. Type E never gives help and has no such marker. The fitnesses of each type are $F(A, x) = (b - c)x$ and $F(E, x) = 0$, respectively, where x is the frequency of type A in the population. Note that A will always have higher fitness, provided $b > c$.

A spiteful greenbeard (S) works in an analogous but opposite way. Type S has an identifiable marker and only harms those who lack the marker. The fitness for a spiteful greenbeard is $F(S, x) = -c(1 - x)$, where x represents the frequency of type S. The fitness of type E (never engages in spite and has no marker) is $F(E, x) = -hx$. Greenbeard spite will be favored whenever $x/(1 - x) > c/h$. Even in this highly idealized model, differences between the evolutionary dynamics of spite and altruism begin to arise. Whether greenbeard spite is favored depends on its frequency in the population, which

is not true of greenbeard altruism. Also, greenbeard spite may be favored even if $c > h$, provided there are enough spiteful greenbeards in the population.

The greenbeard models can be generalized in a useful way. The marker is really just a kind of signal that is perfectly correlated with a behavioral type. Real identifying markers would not be so perfectly aligned. Perhaps the greenbeards of altruists do not develop with perfect reliability, perhaps the ability to correctly identify greenbeard markers is imperfect, or perhaps it is possible to mimic the greenbeard trait in a semireliable way (sometimes called a "falsebeard").

Consider a model of signals and conditional social behavior. Let g represent the success rate of a signal about one's type for eliciting the appropriate conditional social behavior, for example, the reliability of a greenbeard trait for successfully eliciting cooperation from other greenbeards. We later separate the production and recognition of this signal, but for now we simply consider g as the probability of a conditional altruist successfully eliciting help from another conditional altruist (with probably $1 - g$, they are treated as a different type). Likewise, to capture the possibility of mimics, let f represent the success rate of a different type eliciting conditional behavior reserved for similar types. In the altruism case, f is the probability that an egoist elicits help from a conditional altruist. The fitnesses of each type can be expressed as follows, where x is the proportion of A in the population:

$$F(A, x) = xg(b - c) - (1 - x)fc. \tag{1}$$

$$F(E, x) = xfb. \tag{2}$$

We can model the prospects for signal-based conditional spite in an analogous way. Let g represent the probability of a conditionally spiteful individual avoiding spite from another such individual. Let f represent the success rate for an egoist also avoiding spiteful behavior. The fitnesses of each type can be expressed as follows, where x represents the proportion of S:

$$F(S, x) = x(1 - g)(-h - c) - (1 - x)(1 - f)c. \tag{3}$$

$$F(E, x) = x(1 - f)(-h). \tag{4}$$

In both cases, we can derive the conditions under which altruism and spite will be favored: $F(A, x) > F(E, x)$ and $F(S, x) > F(E, x)$, respectively. Conditional altruism will be favored over deceptive egoists when

$$g - f \geq \frac{fc}{x(b - c)}. \tag{5}$$

Conditional spite will be favored over deceptive egoists when

$$g - f \geq \frac{(1 - f)c}{x(h + c)}. \tag{6}$$

The formal similarities between conditions (5) and (6) are apparent. However, framing things in terms of signals that vary in reliability generates important differences. While both conditions require that the signaling types identify one another more reliably than they misidentify mimicking egoists, a highly successful mimic is a more serious threat to conditional altruism than to conditional spite (more on this below). Notice also that the cost-to-harm ratio in (6) plays a different role than the cost-to-benefit ratio in (5).

The differences become most apparent when we consider the stability conditions for each type of greenbeard, which include $F(A, 1) \geq F(E, 1)$ and $F(S, 1) \geq F(E, 1)$ for altruism and spite, respectively. For altruism, this amounts to

$$g - f \geq g\frac{c}{b}. \tag{7}$$

For spite, the condition is

$$g - f \geq (1 - g)\frac{c}{h}. \tag{8}$$

These conditions represent necessary, but not sufficient, conditions for neutral stability. Writing expressions (7) and (8) as strict inequalities would make them sufficient, but not necessary, for evolutionary stability (Maynard Smith 1982).

Provided $g > f$ and $x \approx 1$, a very reliable signal $g \approx 1$ can guarantee that conditional spite has strictly higher fitness than any potential invading type. But, a very high degree of reliability does not necessarily translate to stability for altruism. An effective (but still imperfect) mimic can threaten to invade no matter the success rate of the signal among the altruists. In other words, spite can be stable even if their signal is relatively easy to fake; the same is not true for altruism. Furthermore, if $g \gg f$, it is possible for extremely costly spite to be stable—acts that are more costly to the actor than harmful to the recipient. Extremely costly altruism, however, cannot be stable, for if $c > b$ then equation (7) cannot be satisfied.

The reason for these asymmetries is that in a population of conditionally spiteful types, there is nothing to be gained by attempting to fake a reliable signal. If the signal is reliable, there is very little harming behavior occurring in the population (and likewise, very few individuals paying the cost). Most individuals are already in a best-case scenario of avoiding harm and not paying the cost to harm. In the case of the conditional altruism, however, there is still the temptation to acquire the benefit without paying the cost. Although conditional behavior based on identifiable signals can create both positive and negative assortment, the asymmetry of these cases shows that such a mechanism is more conducive to spite than to altruism.

The same moral is borne out in other more sophisticated models in which similar asymmetries have been noted. For instance, Lehmann, Feldman, and

Rousset (2009) investigate the coevolution of a neutral marker with associated spiteful or altruistic behavior (they use the terms "harming" and "helping"). They find that under certain conditions, harming behavior (spite) is more likely to coevolve with genetic markers than helping behavior (altruism). Interestingly, one of the earliest discoveries of a greenbeard effect involved conditional harming behavior in the red fire ant where workers would kill queens that did not bear the marker (Keller and Ross 1998).

3.1. The Evolution of the Signal. While type-specific signals can stabilize conditional spite and conditional altruism, it remains to be seen whether such signals can coevolve with the social behaviors. Signals require both a sender (e.g., the bearer of a greenbeard marker) and a receiver (e.g., the individuals who may identify the marker and act accordingly; Skyrms 2010). Any investigation to the evolution of signal-mediated conditional social behavior would need to consider both the production of the signal as well as the ability to recognize the signal.

Note that the variables g and f above are not fully specified in this regard. They represent only the probability that the relevant conditional behavior is invoked on the basis of the type sending the signal. Any failure to invoke the corresponding conditional behavior could be due to a failure in the signal (e.g., a beard is not green enough) or a failure in recognition by the other individual (e.g., beard color blindness).

To consider these possibilities, we can introduce a variable representing the reliability of a particular type of conditional strategy producing the signal g_s and the reliability of others with that conditional strategy recognizing the signal g_r. Assuming these are independent, the global g parameter is the product $g = g_s g_r$. For simplicity, we treat the success rate of mimics f as constant.[2]

The fitness functions in the case of conditional altruism and conditional spite become

$$F(A, x) = xg_s g_r b - xg_s g_r c - (1 - x)fc, \tag{9}$$

$$F(S, x) = -x(1 - g_s g_r)h - x(1 - g_s g_r)c - (1 - x)(1 - f)c. \tag{10}$$

We can now consider selection on g_s and g_r. To evaluate how selection may influence g_s, suppose that $g_r = 1$ and that mutants are introduced into the population (A' and S' for the altruism and spite cases, respectively) that differ from the natives only in the reliability of producing their type-specific signal $g_s' \neq g_s$.

2. The models here could be expanded by dividing the f factor in a similar way into f_s and f_r. Doing so complicates the results but does not change the central message, so we have left out a discussion of these results for brevity.

In the case of altruism, the mutant has the fitness

$$F(A', x) = xg_s'b - xg_sc - (1 - x)fc. \tag{11}$$

Under these assumptions, $F(A', x) > F(A, x)$ whenever $g_s' > g_s$. Consequently, we should expect selection to result in more reliable signals among the conditional altruists.

The same is true for conditional spite, where a mutant spiteful individual (S') has the fitness

$$F(S', x) = -x(1 - g_s')h - x(1 - g_s)c - (1 - x)(1 - f)c. \tag{12}$$

In this case, we see an exact parallel with conditional altruism: $F(S', x) > F(S, x)$ whenever $g_s' > g_s$. With both conditional spite and conditional altruism, if recognition is reliable, we should expect the reliability of producing the type-specific signal to increase under selection.

3.2. The Evolution of Recognition. Reliable signal production is only half of the story. We need to also address the ability of others to recognize and respond to the signal. Now suppose that signal production is perfect ($g_s = 1$), but the ability to recognize signals is not ($g_r < 1$). Mutants are introduced who have a conditional strategy that is more or less effective at identifying others with that same strategy ($g_r' \neq g_r$). As before, we assume that the success rate of potential mimic signals is constant.

In the case of altruism, the fitness for the mutant (A' with g_r') is

$$F(A', x) = xg_rb - xg_r'c - (1 - x)fc. \tag{13}$$

Here $F(A', x) > F(A, x)$ if and only if $g_r' < g_r$. That is, any mutant conditional type that is less reliable at identifying others of the same type will have a strict fitness advantage over the natives. Therefore, if conditional altruism does evolve, we should expect the type-recognition ability of this population to erode. Once this erodes, so does the barrier for egoistic invaders.

In the case of spite, the fitness for the mutant (S') is

$$F(S', x) = -x(1 - g_r)h - x(1 - g_r')c - (1 - x)(1 - f)c\,. \tag{14}$$

Notice the difference: when conditional spite is common, any mutant that is more reliable at identifying others of the same type ($g_r' > g_r$) will have a strict fitness advantage over the natives. Thus, if conditional spite evolves, there will be persistent evolutionary pressure for better type recognition.

4. Reciprocity. The previous model considered conditional behavior based on an exogenous signal. The signal influenced the interaction but was not part of the helping or harming behavior. Such a signal is only one way that type

recognition may be mediated. There may be circumstances in which there simply are no indicators of an individual's behavioral type outside of the behavior itself. Yet a limited source of information about behavioral type is still available: observation of past behavior. This ties into another early solution to the evolutionary puzzle of altruism: reciprocity (Trivers 1971; Axelrod and Hamilton 1981). Reciprocity can generate the positive assortment of behaviors necessary for prosocial strategies to evolve (Michod and Sanderson 1985). However, the connection between reciprocity and spite has been largely overlooked. An examination of reciprocal spite reveals that reciprocity can prevent as well as promote prosocial behavior, but there is also reason to think reciprocal spite is less evolutionarily significant than reciprocal altruism.

The theory of repeated games provides a way to capture the potential effects of reciprocity. The general characterization of altruism and spite from above can be represented with games: the Prisoner's Dilemma and the Prisoner's Delight. Imagine that two individuals get the option to confer a benefit b on their partner in the case of the Dilemma (cooperate) or a harm h in the case of the Delight (spite), at some cost c to themselves, and must make their decision simultaneously.

In game theoretic terms, the evolutionary puzzles for altruism and spite are that such strategies are strictly dominated and do not form an equilibrium of their respective games. However, if the game is played more than once, there are strategies that will reciprocate and can stabilize both cooperation and spite in the repeated game—the famous tit for tat is an example of such a strategy in the Dilemma (Axelrod 1984). The Folk theorem shows us that there are an infinite number of possible equilibria in the repeated games (Friedman 1971; Rubinstein 1979). Simply put, the theorem states that any pair of payoffs for which both players are earning more than their minimax payoff can be maintained in equilibrium so long as there is a sufficient probability (p) of a future round of play (see Fudenberg and Tirole 1991).

Applying the Folk theorem to the Dilemma shows that any payoff for which both players receive more than 0 is a potential equilibrium payoff for some pair of appropriately defined strategies and a sufficiently large probability of a future interaction (see fig. 1, left). Note that many possible equilibria are inefficient, when players receive substantially less than the fully cooperative $b - c$ payoffs. There is no guarantee of perfect cooperation in the repeated game, and in fact there are strategies that can extort other players (Press and Dyson 2012); this effectively turns the repeated Dilemma into a bargaining game (Binmore 2005). Nevertheless, compared to the one-shot game, where cooperation cannot be sustained in equilibrium, the repeated Dilemma is much more conducive to cooperation. Reciprocity can only improve the evolutionary prospects for altruism.

The repeated Delight also has solutions that involve the use of the dominated spite strategy. In this case, the minimax payoff is $-h$. This means that

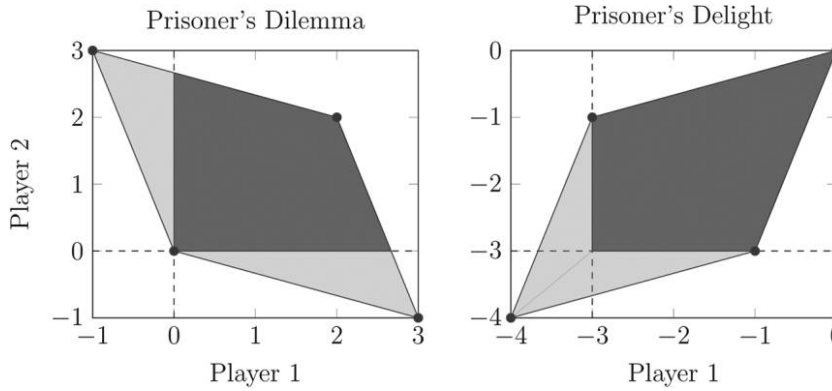

Figure 1. Suppose $b = 3$ and $c = 1$. Shading represents possible mean payoff (per round) for any combination of repeated game strategies. Dark shading represents payoffs for feasible equilibria in the repeated Prisoner's Dilemma (*left*) and Prisoner's Delight (*right*). Color version available as an online enhancement.

there will be many equilibria in the repeated Delight for which individuals receive less than the optimal zero payoff (see fig. 1, right). Spiteful behavior will occur with some regularity in these equilibria.

We can express precisely that altruism can be maintained in equilibrium, provided that the current and expected future payoff outweighs the short-term gain for not cooperating:

$$\sum_{i=1}^{\infty} p^i(b - c) > c. \tag{15}$$

There is no directly analogous condition for the stability of spite because repeated spite is already the worst punishment one individual can inflict on another. Yet spite can be maintained in equilibrium if it occurs only some of the time. Consider an alternating strategy that harms one's partner every Nth round and will harm persistently if the pattern is not reciprocated. This strategy can be maintained in equilibrium if the current and future payoffs outweigh the benefit of avoiding the cost to spite in every Nth round:

$$h\left(\sum_{i=1}^{\infty} p^i - \sum_{i=1}^{\infty} p^{Ni}\right) > c\left(1 + \sum_{i=1}^{\infty} p^{Ni}\right). \tag{16}$$

Notice that with larger values of N and p reciprocity can maintain a limited amount of spite that is more costly to the actor than it is harmful to the recipient. This contrasts with the conditions enabling reciprocal altruism, which require that the benefit to the recipient outweigh the cost to the actor, as well as with the standard inclusive fitness approaches to spite, which require that spite be more harmful to others than it is costly to the actor.

Investigating evolutionary dynamics in repeated games involves restricting the strategy space to a finite subset. Here, we consider the case of an infinite randomly mixing population in which individuals are playing an infinitely repeated game ($p = 1$). Suppose that all pure strategies with a one-round memory are represented and are executed without errors. For a 2×2 symmetric game, there are 32 such strategies: two possible initial strategies and two possible responses for each of the four possible outcomes in the previous round. Also suppose that initial frequencies are drawn from a uniform distribution over the strategy space and that evolution operates according to the discrete-time replicator dynamic (Weibull 1995).

In the context of the Dilemma (with $b = 3$, $c = 1$, and $p = 1$), the conditions for stability of cooperative equilibria are met. Indeed, the typical result of evolution is some combination of altruistic strategies (strategies that result in a perfect cooperation rate when played among one another). Cooperation evolved in every simulated population.[3] In the repeated Delight (with $h = 3$, $c = 1$, and $p = 1$), reciprocal spite is possible with strategies that alternate between harming and not harming one's partner.[4] In this case, the conditions for the stability of reciprocal spite are very similar to those of altruism. Nevertheless, reciprocal spite evolves very rarely: only five in 10^5 simulated populations reached an equilibrium with reciprocal spite. In each case, the populations involved a collection of alternating strategies.

Reciprocal spite and reciprocal altruism are both evolutionarily possible but with important differences. Reciprocal spite allows for the possibility of extremely costly harming behavior, where the cost of an individual spiteful act outweighs the harm done. Reciprocal altruism does not allow for extremely costly helping behavior. Also, in limited models with restricted strategy spaces, reciprocal altruism evolves far more readily than reciprocal spite.

It is also worth noting that reciprocity can be indirect when behaviors are enforced by others not directly affected by transgressions. Alexander (1987) argued that indirect reciprocity is the crucial feature of human moral systems. Further, it has been shown that if there are reliable ways of tracking past behavior, conditional strategies can stabilize altruism by indirect reciprocity (Nowak and Sigmund 2005), and similar mechanisms can support the evolution of spite (Johnstone and Bshary 2004). Interestingly, in models where the information transmission coevolves with social behavior, it has been observed that cooperation through indirect reciprocity faces evolutionary obstacles due to subversion of the information transmission (Smead 2010).

3. Simulations were written in C, and 10^5 independent trials were run.

4. Fudenberg and Maskin (1990) show that this kind of alternating strategy is evolutionarily significant in the context of the repeated Prisoner's Dilemma.

5. Recognition and Morality. Altruism and spite have a common evolutionary core: assortment. In both cases the requisite assortment can come about through recognition and conditional social behavior. Despite this commonality, the prospects for the coevolution of the specific mechanisms of assortment stand in stark contrast for spite and altruism. Our models suggest that assortment driven by recognition via exogenous signals can coevolve with spite but not with altruism. The inverse is true for assortment driven by reciprocity. Thus, both of these social behaviors may play different but essential roles in the evolution of moral norms more generally.

The models here may be particularly relevant to understanding the evolution of punishment. One standard account presumes that cooperation evolves first among kin or small groups, usually involving mutually beneficial social interactions, then the scope of cooperative behavior expands to include larger number of individuals with the help of punishment to stabilize cooperation against the increasing risks of free riding or subversion (Axelrod and Hamilton 1981; Boyd et al. 2003; Binmore 2005; Sterelny 2012). Punishment requires recognition of individuals and their behavioral tendencies for such sanctions to maintain cooperative norms. Insofar as punishment is a deep feature of human moral systems, understanding the evolutionary origins of recognition will be a key component of a robust account of the evolution of moral norms. The results here illustrate important evolutionary connections behind the manner in which recognition occurs and the associated social interactions. For instance, what kinds of social interactions can provide the necessary scaffolding for the evolution of recognition? The answer to this question depends on the type of information being used in the recognition process. If it is direct information of past behaviors (as occurs in reciprocity), altruistic interactions can support the evolution of recognition. But, if the recognition is reliant on exogenous signals, altruistic interactions lead to a deterioration in recognition ability. Conditional spite, however, can scaffold recognition when exogenous signals matter.

Perhaps even more striking, the models we have explored here and elsewhere show that conditional harming behavior can even evolve independently from its potential use of as punishment to enforce prosocial norms. In contrast to the standard account, the conditional harming seen in punishment may have evolved first, scaffolding the evolution of recognition and thereby enabling robust cooperation to evolve later.

Nietzsche made a conjecture about the origin of punishment. After defending the methodological point that historical inquiry must respect that "the cause of the origin of a thing and its eventual utility, its actual employment and place in a system of purposes, lie worlds apart" (1887/1967, II 12)—a point Gould and Lewontin (1979) famously echo in the context of evolutionary inquiry—Nietzsche claims: "In accordance with the [stated] major point of historical method, it is assumed without further ado that the procedure

[of inflicting harm] itself will be something older, earlier than its employment in punishment, that the latter is only *projected* and interpreted *into* the procedure (which had long existed but been employed in another sense)" (1887/1967, II 13). Nietzsche may very well be right.

REFERENCES

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

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

Axelrod, R., and W. D. Hamilton. 1981. "The Evolution of Cooperation." *Science* 211:1390–96.

Binmore, K. 2005. *Natural Justice*. Oxford: Oxford University Press.

Boyd, R., H. Gintis, S. Bowles, and P. J. Richerson. 2003. "The Evolution of Altruistic Punishment." *Proceedings of the National Academy of Sciences USA* 100:3531–35.

Dawkins, R. 1976. *The Selfish Gene*. Oxford: Oxford University Press.

Fletcher, J. A., and M. Doebeli. 2009. "A Simple and General Explanation for the Evolution of Altruism." *Proceedings of the Royal Society of London* B 276:13–19.

Friedman, J. W. 1971. "A Non-cooperative Equilibrium for Supergames." *Review of Economic Studies* 38 (1): 1–12.

Fudenberg, D., and E. Maskin. 1990. "Evolution and Cooperation in Noisy Repeated Games." *American Economic Review* 80:274–79.

Fudenberg, D., and J. Tirole. 1991. *Game Theory*. Cambridge, MA: MIT Press.

Gardner, A., and S. A. West. 2009. "Greenbeards." *Evolution* 64:25–38.

Gould, S. J., and R. C. Lewontin. 1979. "The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme." *Proceedings of the Royal Society of London* B 205:581–98.

Hamilton, W. D. 1964. "The Genetical Evolution of Social Behaviour." Pt. 1. *Journal of Theoretical Biology* 7:1–16.

———. 1975. "Innate Social Aptitudes of Man: An Approach from Evolutionary Genetics." In *Biosocial Anthropology*, ed. R. Fox. London: Malaby.

Johnstone, R. A., and R. Bshary. 2004. "Evolution of Spite through Indirect Reciprocity." *Proceedings of the Royal Society of London* B 271:1917–22.

Keller, L., and K. G. Ross. 1998. "Selfish Genes: A Green Beard in the Red Fire Ant." *Nature* 394:573–75.

Lehmann, L., K. Bargum, and M. Reuter. 2006. "An Evolutionary Analysis of the Relationship between Spite and Altruism." *Journal of Evolutionary Biology* 19:1507–16.

Lehmann, L., M. W. Feldman, and F. Rousset. 2009. "On the Evolution of Harming and Recognition in Finite Panmictic and Infinite Structured Populations." *Evolution* 63:2896–2913.

Maynard Smith, J. 1982. *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.

Michod, R. E., and M. Sanderson. 1985. "Behavioural Structure and the Evolution of Social Behaviour." In *Evolution: Essays in Honour of John Maynard Smith*. Cambridge: Cambridge University Press.

Nietzsche, F. 1887/1967. *On the Genealogy of Morals*. Ed. W. Kaufmann and R. J. Hollingdale. New York: Vintage.

Nowak, M. A., and R. M. May. 1992. "Evolutionary Games and Spatial Chaos." *Nature* 359:826–29.

Nowak, M. A., and K. Sigmund. 2005. "Evolution of Indirect Reciprocity." *Nature* 437:1291–98.

Pollack, G. B. 1989. "Evolutionary Stability on a Viscous Lattice." *Social Networks* 11:175–212.

Press, W. H., and F. J. Dyson. 2012. "Iterated Prisoner's Dilemma Contains Strategies That Dominate Any Evolutionary Opponent." *Proceedings of the National Academy of Sciences USA* 109:10409–13.

Price, G. R. 1970. "Selection and Covariance." *Nature* 277:520–21.

Rubinstein, A. 1979. "Equilibrium in Supergames with the Overtaking Criterion." *Journal of Economic Theory* 21:1–9.

Skyrms, B. 1996. *Evolution of the Social Contract*. Cambridge: Cambridge University Press.
———. 2010. *Signals: Evolution, Learning, and the Flow of Information*. Oxford: Oxford University Press.
Smead, R. 2010. "Indirect Reciprocity and the Evolution of 'Moral Signals.'" *Biology and Philosophy* 25:33–51.
Smead, R. S., and P. Forber. 2013. "The Evolutionary Dynamics of Spite in Finite Populations." *Evolution* 67:698–707.
Sober, E., and D. S. Wilson. 1998. *Unto Others*. Cambridge, MA: Harvard University Press.
Sterelny, K. 2012. *The Evolved Apprentice*. Cambridge, MA: MIT Press.
Tomasello, M., and A. Vaish. 2013. "Origins of Human Cooperation and Morality." *Annual Review of Psychology* 64:231–55.
Trivers, R. L. 1971. "The Evolution of Reciprocal Altruism." *Quarterly Review of Biology* 46:35–57.
Weibull, J. W. 1995. *Evolutionary Game Theory*. Cambridge, MA: MIT Press.
West, S. A., and A. Gardner. 2010. "Altruism, Spite, and Greenbeards." *Science* 327:1341–44.