Selfishness as second-order altruism

Omar Tonsi Eldakar*† and David Sloan Wilson*‡

Departments of *Biological Sciences and ‡Anthropology, Binghamton University, Binghamton, NY 13902-6000

Edited by Brian Skyrms, University of California, Irvine, CA, and approved March 6, 2008 (received for review December 22, 2007)

Selfishness is seldom considered a group-beneficial strategy. In the typical evolutionary formulation, altruism benefits the group, selfishness undermines altruism, and the purpose of the model is to identify mechanisms, such as kinship or reciprocity, that enable altruism to evolve. Recent models have explored punishment as an important mechanism favoring the evolution of altruism, but punishment can be costly to the punisher, making it a form of second-order altruism. This model identifies a strategy called ''selfish punisher'' that involves behaving selfishly in first-order interactions and altruistically in second-order interactions by punishing other selfish individuals. Selfish punishers cause selfishness to be a self-limiting strategy, enabling altruists to coexist in a stable equilibrium. This polymorphism can be regarded as a division of labor, or mutualism, in which the benefits obtained by first-order selfishness help to ''pay'' for second-order altruism.

punishment | cooperation | mutualism | game theory | public goods

Selfishness is rarely described as a group-beneficial strategy. Selfish strategies are labeled as deviant, cheating, free-riding, egoistic (1), but most of all, as undermining altruism and cooperation (2). In contrast, altruistic and cooperative strategies, almost by definition, benefit the group, often at the expense of the individual actor (2). In the typical evolutionary model, the invasion of selfish strategies into a group leads to the dissolution of altruism. Examples include scroungers among foraging groups (3, 4), infanticide of unrelated infants (5), sneaking worker reproduction in eusocial insect colonies (6, 7), and failure to help in territorial defense (8, 9). The experimental economics literature amply demonstrates the corrosive effects of selfishness in human social interactions. In public-goods games, participants start out moderately generous but quickly withdraw their cooperation in the presence of selfish cheaters (10–13).

Earlier evolutionary models focused on how altruism can evolve through nonrandom interactions or guarded cooperation in dyadic interactions (14–17). More recently, interest has focused on punishment as a mechanism for maintaining altruism in sizeable groups (10, 11, 13, 18–23). Punishment can be effective in curtailing selfish behavior within a group, but it can also be costly for the punisher, compared with cooperators in the same group who do not punish, thereby qualifying as a form of second-order altruism. Individuals who are altruistic in first-order and second-order interactions are at a double disadvantage (10, 23–25). The solution to this problem might lie where least expected.

An often overlooked aspect of game theory is that selfish individuals also have an incentive to punish other selfish individuals, thereby increasing the proportion of cooperators for them to exploit. This behavior might seem hypocritical in moral terms, but it makes sense as an evolutionary strategy. It can even be looked upon as a division of labor, or mutualism, whereby cheating during first-order interactions becomes a ''payment'' for altruism (punishment) in second-order interactions. A combination of strategies (selfish punisher plus altruistic nonpunisher) that split the costs of first- and second-order altruism can be superior to a single-altruist/punisher strategy that bears both costs.

In an earlier computer simulation model (26, 27), we showed that when first- and second-order altruism are modeled as initially uncorrelated traits, a negative correlation robustly develops between the two, although the size of the correlation depends upon a number of parameters. Here, we present an analytical model demonstrating how altruistic nonpunishers and selfish punishers, through the benefit of division of labor, can exist in a stable equilibrium.

The Model

First-Order Altruism and Selfishness. The model emulates an experimental economics game in which each member of a group is provided an endowment, b, that can be kept or invested in a public good. The combined investment in the public good is multiplied by a factor, m, and distributed equally to everyone in the group. The total payoff of each individual (the proportion of the endowment kept for oneself plus one's share of the public good) is assumed to be linearly related to fitness. This scenario can easily be related to biological situations, such as investing effort in a hunt in which everything captured will be shared. The model considers the two pure strategies of investing all (altruist) or none (selfish) of one's endowment. For a model that includes a continuum of endowment strategies, see ref. 26. If p is the proportion of altruists within a single group, then the increments in the fitness of the altruists and selfish types after one round of interactions is:

$$W_A = pbm \quad [1]$$

$$W_S = pbm + b \quad [2]$$

Baseline fitness is ignored because it is the same for both types. It is clear that in the absence of punishment, selfish individuals always have the highest fitness within the group because all group members obtain an equal share of the contributions from the altruists, yet selfish individuals keep rather than donate b. Therefore, an altruist would obtain a greater fitness by switching to the selfish noncontributing alternative, resulting in a net gain of

$$b - \left(p - \frac{1}{N}\right)bm$$

(selfish gain of b minus the now share of the reduced group payoff caused by the loss of a single altruist). However, the group has the highest fitness when everyone is an altruist, resulting in the classic prisoner's dilemma situation.

Punishment. Now consider a method of social control in which individuals can punish selfish group members at a personal cost c, which results in the selfish individual losing its acquired energy for a given round of the game. The cost is incurred for every selfish member of the group, and punished individuals risk the loss of their previously acquired energy ($b + pbm$) at a proba-

Author contributions: O.T.E. and D.S.W. designed research; O.T.E. performed research; and O.T.E. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

†To whom correspondence should be addressed. E-mail: oeldakar@gmail.com.

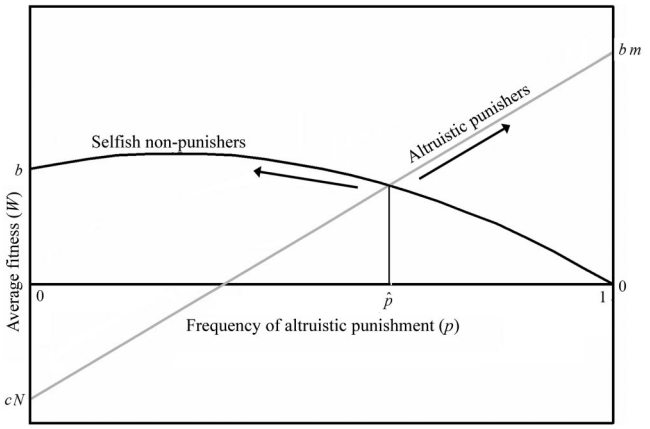

Fig. 1. Fitness plot of the average within-group fitness of altruistic punishers at selfish nonpunishers at varying frequencies of altruistic punishers. Arrows indicate the population trajectories at any given frequency. Altruistic punishers maintain a fitness advantage and evolve to fixation when beyond the bistable equilibrium point ($\hat{p}$) and visa versa for selfish nonpunishers.

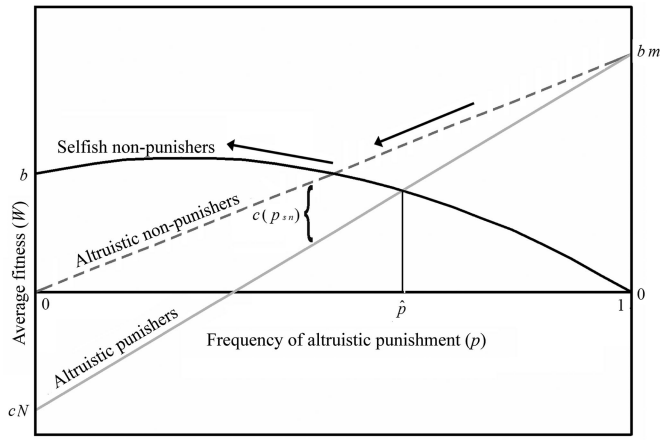

Fig. 2. Fitness plot of the average within-group fitness of altruistic punishers, selfish nonpunishers, and the invading altruistic nonpunishers strategy at varying frequencies of altruistic punishers. At invasion at a frequency >0, the fitness of altruistic nonpunishers resides above that of altruistic punishers by the value of the cost of punishment $c(p_{sn})$. Arrows indicating the population trajectories illustrate the failure of altruistic punishers to maintain the highest fitness in their previous region of selective advantage, resulting in their decrease in frequency and the establishment of selfish nonpunishers as an ESS.

bility equal to the frequency of punishers that reside in the group. Punishment can be exhibited by either altruists or selfish individuals, yielding four possible strategies (altruistic punishers, selfish punishers, altruistic nonpunishers, selfish nonpunishers) at frequencies of p_{ap}, p_{sp}, p_{an}, and p_{sn}, respectively. In a single group, the average fitness W of the four strategies are as follows:

$$W_{an} = (p_{an} + p_{ap})bm \quad [3]$$

The fitness of altruistic nonpunishers (Eq. **3**) is exclusively based on the altruistic contributions of group members and remains unaffected by punishment.

$$W_{sn} = [(p_{an} + p_{ap})bm + b](1 - p_{ap} - p_{sp}) \quad [4]$$

The fitness of selfish nonpunishers (Eq. **4**) is determined by the sum of withholding cooperation b, and the exploitation of altruists $(p_{an} + p_{ap})bm$, however, is only retained in the proportion that punishers are not present in the group $[1 - (p_{ap} + p_{sp})]$.

$$W_{ap} = (p_{an} + p_{ap})bm - c(p_{sn} + p_{sp}) \quad [5]$$

Altruistic punishers (Eq. **5**) retain the same fitness as altruistic nonpunishers yet bear a cost for punishing all selfish members within the group.

$$W_{sp} = [(p_{an} + p_{ap})bm + b]\left[1 - \left(p_{ap} + p_{sp} - \frac{1}{N}\right)\right] - c\left(p_{sn} + p_{sp} - \frac{1}{N}\right) \quad [6]$$

Selfish punishers (Eq. **6**) exploit the altruistic contributions of group members yet punish all other selfish group members.

Punishment reduces the fitness of selfish strategies, lowering their fitness within groups. However, punishers in turn have a lower fitness than nonpunishers within the same group. The act of punishment, taken by itself, can be regarded as altruistic because punishers pay a personal cost to provide a benefit (the removal of selfish individuals) that is shared by nonpunishers. We will explore the dynamics of the model by first considering two of the strategies and then adding additional strategies.

Altruistic Punishers vs. Selfish Nonpunishers. Altruistic punishers and selfish nonpunishers form an unstable equilibrium with each strategy exhibiting positive frequency dependence, as shown in Fig. 1. Altruistic punishment can evolve if it begins above a threshold frequency. However, its stability is eliminated with the addition of altruistic nonpunishers, as shown in Fig. 2. By not partaking in the costs of punishment, altruistic nonpunishers outcompete altruistic punishers and undermine punishment, resulting in the establishment of selfishness as an ESS (for population trajectories, see Fig. 3). To summarize the model so far, the viability of altruistic punishment requires two conditions: their initial frequency must be sufficiently high, and second-

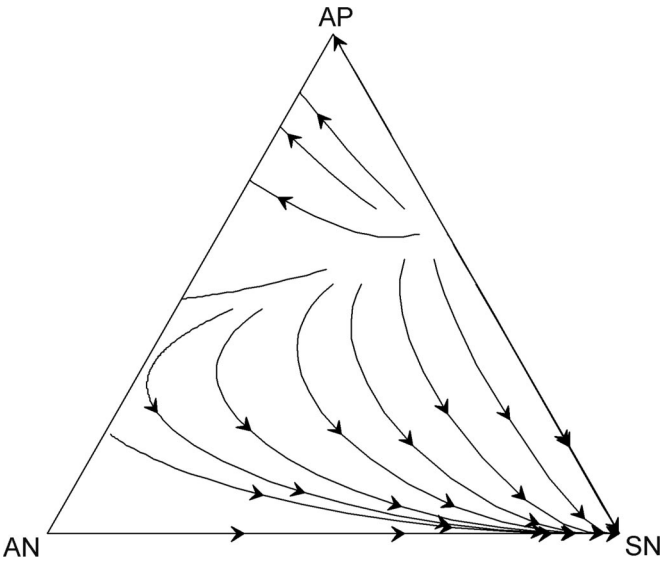

Fig. 3. Triangle plot illustrating the population trajectories for altruistic nonpunishers (AN), altruistic punishers (AP), and selfish nonpunishers (SN) at all possible initial frequencies. Each corner represents the population as that pure strategy. Both altruistic punisher and selfish nonpunisher strategies exhibit positive density dependence, attracting populations to their respective corners. Under conditions of AP advantage, populations will approximate the AP–AN axis, and because of the absence of the SN strategy both strategies will exhibit equal fitness. The population will traverse the AN–AP axis via drift, allowing the population to drop below the region of AP attraction and become vulnerable to SN invasion and subsequent fixation.

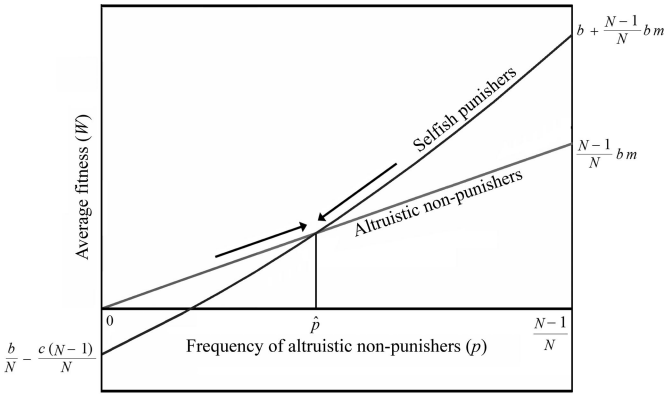

Fig. 4. Fitness plot of the average within-group fitness of selfish punishers and altruistic nonpunishers. Arrows indicate the population trajectories at any given frequency. Both altruistic nonpunishers and selfish punishers exhibit negative frequency dependence. Each strategy maintains a greater fitness when present at low frequencies and loses the fitness advantage as that strategy becomes more common, causing the population to settle at the stable equilibrium point ($\hat{p}$).

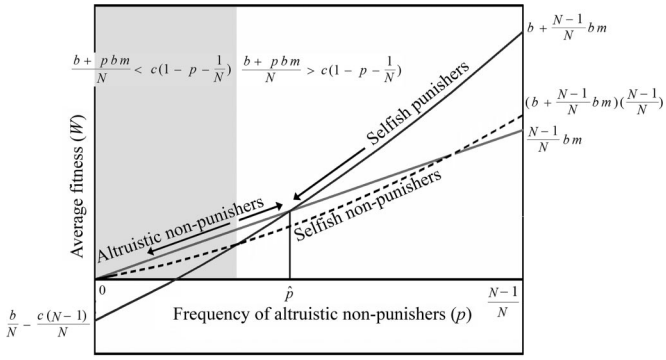

Fig. 5. Fitness plot of the average within-group fitness of selfish punishers, altruistic nonpunishers, and the invading selfish nonpunishers strategy at varying frequencies of altruistic nonpunishers. Arrows indicating the population trajectories illustrate that the selfish punisher and altruistic nonpunisher equilibrium point ($\hat{p}$) remains resistant to the invasion of selfish nonpunishers given $b + p_{an}bm/N > c(1 - p_{an} - 1/N)$.

order free-riding (altruistic nonpunishment) must be excluded as a superior strategy.

To curtail second-order free-riding, punishment can be expanded to include altruists who do not punish. This tactic is vulnerable to third-order free-riding, leading to a seemingly infinite regress (for a possible solution to this problem, see ref. 28). Alternatively, punishers can recoup their cost of punishment by being selfish in first-order interactions.

Selfish Punishment. Selfish punishment is a self-limiting strategy that provides a form of second-order altruism by punishing selfishness. In the course of punishing all selfish individuals (including other selfish punishers), selfish punishers remain restricted through negative density dependence, promoting altruism within groups. In a simple two-strategy model, selfish punishers and altruistic nonpunishers form a stable equilibrium, as shown in Fig. 4. Under certain conditions, this equilibrium is robust against the invasion of selfish nonpunishers.

$$\frac{b + p_{an}bm}{N} > c\left(1 - p_{an} - \frac{1}{N}\right)$$

Because selfish punishers refrain from expelling themselves from the group, an individual punisher has a

$$N - \frac{1}{N}$$

less chance of expulsion than a selfish nonpunisher within the same group. Therefore, if the benefit of remaining in the group

$$\frac{b + p_{an}bm}{N}$$

is greater than the cost of punishing all other selfish individuals

$$c\left(1 - p_{sn} - \frac{1}{N}\right),$$

the equilibrium between selfish punishers and altruistic nonpunishers remains stable. This equilibrium range is expanded with increasing the values of b and m as well as increasing the frequency of altruists residing in the group. Furthermore, decreasing group size N results in a greater discrepancy between the chances of expulsion, favoring selfish punishers over selfish nonpunishers within a group.

Groups comprising selfish punishers and altruistic nonpunishers will reconstitute the stable equilibrium regardless of initial frequencies, provided selfish nonpunishers are sufficiently scarce in the population (Fig. 5; for population trajectories, see Fig. 6). Once established, the within-group stability will maintain altruism at a constant frequency in a multigroup population.

Discussion

The majority of evolutionary models of social behavior have treated selfishness as an impediment to altruism (14, 16, 29, 30). When punishment is modeled as a mechanism that facilitates the evolution of altruism, the punishers are usually assumed to be altruists (13, 18, 19–23). Additional mechanisms that might facilitate the evolution of altruism include conformance transmission within groups (31) and withholding cooperation from

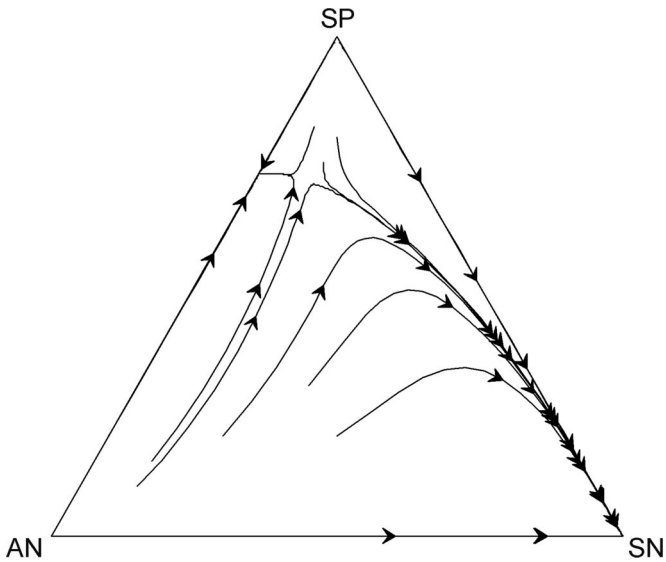

Fig. 6. Triangle plot illustrating the population trajectories for altruistic nonpunishers (AN), selfish punishers (SP), and selfish nonpunishers (SN) at all possible initial frequencies. Each corner represents the population as that pure strategy. If selfish nonpunishers are sufficiently scarce, the population will settle at the selfish punisher–altruistic nonpunisher equilibrium point. Under conditions with a greater frequency of selfish nonpunishers, the population will be attracted to the selfish nonpunisher corner.

selfish individuals (32). In all of these models, the possibility that selfishness might facilitate the evolution of altruism has been largely overlooked.

Inherent in evolutionary game theory is reduction in the fitness of selfish individuals with the increasing frequency of selfishness within groups, which implies that selfish individuals, in addition to altruists, have an incentive to punish selfishness (33). Selfish individuals are in a better position to do so because they have increased their fitness by their first-order interactions. Altruistic punishment is the least-fit strategy within a single group because of the double cost of first- and second-order altruism.

Here, we have shown that selfish strategies during first-order interactions can provide a more stable form of punishment within groups than altruistic punishers. Although this model is constrained to one round of interactions, a multiround computer simulation model with a continuum of altruism and punishment strategies (26) reaches similar conclusions, resulting in a high frequency of altruistic nonpunishers and a low frequency of selfish punishers that keep selfishness at bay. Consistent with predictions from the analytic model presented here, as punishment costs rise or group size is increased in the simulation model, punishers contribute less toward cooperation in the simulation model, exemplifying the tradeoff between first- and second-order altruism. Interestingly, selfish punishers do not evolve to overexploit altruistic nonpunishers beyond retrieving the costs of punishment.

A discussion of terminology is in order because the definition of the terms "altruism" and "selfishness" are notoriously variable (2, 34). For example, altruism is defined in terms of relative fitness within groups in multilevel selection models and in terms of absolute fitness in inclusive fitness theory models (35). When an individual increases the absolute fitness of a recipient at an absolute cost to itself, its behavior appears unambiguously altruistic. Yet, when the same individual is paired with another altruist, the trading of benefits makes the behavior appear like a form of self-interested mutualism. Selfishness can be individually advantageous in a group without punishers but maladaptive in a group with punishers, and so on.

We do not insist upon a given set of definitions (for a discussion of pluralism in evolutionary models of social behavior, see refs. 36–38), as long as sufficient information is provided to translate among perspectives. We have based our definitions on what takes place during each stage of a two-stage interaction. Thus, contributing to a public good counts as altruistic, and keeping one's endowment counts as selfish during the first stage, even though punishment during the second stage might cause selfishness during the first stage to become maladaptive. Similarly, punishment during the second stage counts as altruistic, and failing to punish counts as selfish, even though the combined two-stage strategy of selfish punishment can be selectively advantageous within groups, given a sufficient number of altruistic nonpunishers. These definitions enable us to retain the same definitions for behaviors during a single stage, as opposed to changing definitions based on whether the behaviors are advantageous or disadvantageous in various combinations. However, we acknowledged that the combination of selfish punishers and altruistic nonpunishers can be viewed as an internally stable mutualism and therefore self-interested in that sense of the word.

In most models of altruism, including altruistic punishment, the altruistic behavior is selectively disadvantageous within groups and requires the differential productivity of groups (between-group selection) to be maintained in the population. The outcome depends upon the balance between levels of selection. A trait that has a large effect on collective fitness at a very small personal cost can easily evolve by group selection (e.g., in randomly formed sizeable groups), despite qualifying as altruistic, as the term is defined within multilevel selection theory.

Some evolutionary models of social behavior result in multiple local equilibria, which are internally stable by definition but can differentially contribute to the total gene pool. Group selection among local equilibria can result in social interactions that are simultaneously good for the group and selectively advantageous within groups. Conformance transmission can cause punishment to become internally stable within groups (31). Our model provides another way for punishment to become internally stable, enabling groups with punishers to persist indefinitely and out-compete groups without punishers over the long term. Group selection might also favor the evolution of modifier traits that adjust the internal equilibrium so that most individuals are altruistic nonpunishers with only a few selfish punishers.

In nonhuman species, selfish punishment (which was termed "corrupt policing") has been observed in the tree wasp *Dolichovespula sylvestris*. Workers that police the reproduction of other workers are more likely to lay eggs themselves (7). Worker policing in monogynous and monandrous colonies of the wasp *Polistes chinensis antennalis* exceeds that predicted by relatedness and was partly attributed to the immediate reproductive benefit of selfishly laid eggs (39). Scrub jays that tend to steal caches from other jays are also more defensive of their own caches, providing evidence of selfish behavior as a self-limiting strategy (40).

Game theorists refer to a "replicator dynamic" as any process whereby the most successful behavioral strategy increases in frequency in a population, which includes but goes beyond genetic evolution (41). The concept of selfish punishment therefore might apply to a diversity of human social systems, from the Castellians of medieval Europe (42), to mafia-like protection rackets, to the formal compensation of specialized punishers (police) for their services. In psychological research, a study using fictional scenarios to investigate altruistic punishment revealed that individuals who were most inclined to cheat were also most inclined to punish and spent more to punish, in part because they had amassed more as a result of their selfish behavior (33). We hope that our model stimulates interest in the concept of selfish punishment in both humans and nonhuman species.

ACKNOWLEDGMENTS. We are indebted to D. L. Farrell, A. B. Clark, J. Shepherd, P. Richerson, K. Panchanathan, M. Dlugos, D. O'Brien, The E. N. Huyck Preserve and members of EvoS, Binghamton University Evolutionary Studies Program, for helpful discussion. This work was supported by the National Science Foundation Alliance for Graduate Education and the Professoriate.

1. Becker GS (1976) Altruism, egoism and genetic fitness: Economics and sociobiology. *J Econ Lit* 14:817–826.
2. Wilson DS, Wilson EO (2007) Rethinking the theoretical foundation of sociobiology. *Q Rev Biol* 82:327–348.
3. Barnard CJ, Sibly RM (1981) Producers and scroungers: A general model and its application to captive flocks of house sparrows. *Anim Behav* 29:343–550.
4. Beauchamp G, Giraldeau LA (1996) Group foraging revisited: Information-sharing or producer-scrounger game? *Am Nat* 147:738–743.
5. Packer C, Pusey AE (1983) Adaptations of female lions to infanticidal incoming males. *Am Nat* 121:716–728.
6. Monnin T, Ratnieks FLW (2001) Policing in queenless ponerine ants. *Behav Ecol* 50:97–108.
7. Wenseleers T, Tofilski A, Ratnieks FLW (2005) Queen and worker policing in the tree wasp *Dolichovespula sylvestris. Behav Ecol Sociobiol* 58:80–86.
8. Grinnell J (2002) Modes of cooperation during territorial defense by African lions. *Hum Nat Int Bios* 13:85–104.
9. Grinnell J, Packer C, Pusey AE (1995) Cooperation in male lions: Kinship, reciprocity or mutualism? *Anim Behav* 49:95–105.
10. Yamagishi T (1986) The provision of a sanctioning system as a public good. *J Pers Soc Psychol* 51:110–116.
11. Yamagishi T (1988) Seriousness of social dilemmas and the provision of a sanctioning system. *Soc Psych Q* 51:32–42.
12. Fehr E, Gachter S (2000) Cooperation and punishment in public goods experiments. *Am Econ Rev* 90:980–994.
13. Fehr E, Gachter S (2002) Altruistic punishment in humans. *Nature* 415: 137–140.
14. Hamilton WD (1964) The genetical evolution of social behaviour, I and II. *J Theor Biol* 7:1–16.

15. Hamilton WD (1975) *Biosocial Anthropology*, ed Fox R (Malaby Press, London), pp 133–155.
16. Axelrod R, Hamilton WD (1981) The evolution of cooperation. *Science* 211:1390–1396.
17. Maynard Smith J (1982) *Evolution and the Theory of Games* (Cambridge Univ Press, Cambridge, UK).
18. Gintis H (2000) Strong reciprocity and human sociality. *J Theor Biol* 206:169–179.
19. Fehr E, Fischbacher U, Gachter S (2002) Strong reciprocity, human cooperation, and the enforcement of social norms. *Hum Nat Int Bios* 13:1–25.
20. Boyd R, Gintis H, Bowles S, Richerson PJ (2003) The evolution of altruistic punishment. *Proc Natl Acad Sci USA* 100:3531–3535.
21. Fehr E, Fischbacher U (2003) The nature of human altruism. *Nature* 425:785–791.
22. Gintis H, Bowles S, Boyd R, Fehr E (2003) Explaining altruistic behavior in humans. *Evol Hum Behav* 24:153–172.
23. Fehr E (2004) Don't lose your reputation. *Nature* 432:449–450.
24. Bowles S, Gintis H (2002) Homo reciprocans. *Nature* 415:125–128.
25. Bowles S, Gintis H (2004) The evolution of strong reciprocity: Cooperation in heterogeneous populations. *J Theor Popul Biol* 65:17–28.
26. Eldakar OT, Farrell DL, Wilson DS (2007) Selfish punishment: Altruism can be maintained by competition among cheaters. *J Theor Biol* 249:198–205.
27. Nakamaru M, Iwasa Y (2006) The coevolution of altruism and punishment: Role of the selfish punisher. *J Theor Biol* 240:475–488.
28. Boyd R, Richerson PJ (1992) Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol Sociobiol* 13:171–195.
29. Sober E, Wilson DS (1998) *Unto Others*: *The Evolution and Psychology of Unselfish Behavior* (Harvard Univ Press, Cambridge, MA).
30. Trivers RL (1971) The evolution of reciprocal altruism. *Q Rev Biol* 46:35–57.
31. Henrich J, Boyd R (2001) Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *J Theor Biol* 208:79–89.
32. Panchanathan K, Boyd R (2004) Indirect reciprocity can stabilize cooperation without the second-order free-riding problem. *Nature* 432:499–502.
33. Eldakar OT, Wilson DS, O'Gorman R (2006) Emotions and actions associated with altruistic helping and punishment. *Evol Psychol* 4:274–286.
34. Wilson DS, Dugatkin LA (1992) Altruism. *Key Words in Evolutionary Biology*, eds Keller EF, Lloyd EA (Harvard Univ Press, Cambridge, MA), pp 29–33.
35. Wilson DS (2004) What is wrong with absolute individual fitness? *Trends Ecol Evol* 19:245–248.
36. West SA, Griffin AS, Gardner A (2007) Social semantics: How useful has group selection been? *J Evol Biol* 21:405–420.
37. West SA, Griffin AS, Gardner A (2008) Altruism, cooperation, mutualism, strong reciprocity, and group selection. *J Evol Biol* 20:415–432.
38. Wilson DS (2008) Social semantics: Toward a genuine pluralism in the study of social behaviour. *J Evol Biol* 21:368–373.
39. Saigo T, Tsuchida K (2004) Queen and worker policing in monogynous and monandrous colonies of a primitively eusocial wasp. *Proc R Soc London Ser B* 271:S509–S512.
40. Emery NJ, Clayton NS (2001) Effects of experience and social context on prospective caching strategies by scrub jays. *Nature* 414:443–446.
41. Gintis H (2000) *Game Theory Evolving* (Princeton Univ Press, Princeton, NJ).
42. Bisson TN (1994) The feudal revolution. *Past Present* 142:6–42.