vol. 164, no. 6 the american naturalist december 2004

Cooperation and Punishment, Especially in Humans

Andy Gardner[*] **and Stuart A. West**[†]

Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, United Kingdom

Submitted January 8, 2004; Accepted August 4, 2004; Electronically published October 18, 2004

abstract: Explaining altruistic cooperation is one of the greatest challenges faced by sociologists, economists, and evolutionary biologists. The problem is determining why an individual would carry out a costly behavior that benefits another. Possible solutions to this problem include kinship, repeated interactions, and policing. Another solution that has recently received much attention is the threat of punishment. However, punishing behavior is often costly for the punisher, and so it is not immediately clear how costly punishment could evolve. We use a direct (neighbor-modulated) fitness approach to analyze when punishment is favored. This methodology reveals that, contrary to previous suggestions, relatedness between interacting individuals is not crucial to explaining cooperation through punishment. In fact, increasing relatedness directly disfavors punishing behavior. Instead, the crucial factor is a positive correlation between the punishment strategy of an individual and the cooperation it receives. This could arise in several ways, such as when facultative adjustment of behavior leads individuals to cooperate more when interacting with individuals who are more likely to punish. More generally, our results provide a clear example of how the fundamental factor driving the evolution of social traits is a correlation between social partners and how this can arise for reasons other than genealogical kinship.

Keywords: kin selection, neighbor-modulated fitness, repression of competition, public-goods game, human evolution, policing.

Explaining cooperation at all levels of biological complexity remains one of the greatest problems for evolutionary biology (Hamilton 1964; Buss 1987; Maynard Smith and Szathmáry 1995). The question is, Why would an individual perform a costly altruistic behavior that benefits another individual? The solutions to this problem that have attracted the most attention are when social partners are related (kin selection, in a general sense; Hamilton 1963, 1964, 1970) or when there is some mechanism for repressing competition between groups (see table 1), such as through repeated interactions/reputation (reciprocity; Trivers 1971; Alexander 1979, 1987; Frank 2003), policing (Ratnieks 1988; Frank 1995, 2003), and systems of rewards or punishments (Oliver 1980; Sigmund et al. 2001). The fundamental similarity between all these mechanisms is that they involve positive correlations between the behaviors played by social partners, which are crucial for the evolution of social behaviors (Hamilton 1975; Grafen 1985; Nee 1989; Frank 1998; Woodcock and Heath 2002).

Here, we are concerned with whether and how punishment can favor cooperation and how this translates into a selective benefit for punishers. The possible role of punishment has recently attracted much theoretical attention, especially with respect to its possible role in favoring cooperation among humans (Hirshleifer and Rasmusen 1989; Boyd and Richerson 1992; Sober and Wilson 1998; Sell and Wilson 1999; Fehr and Gächter 2000). However, the mechanism underlying these previous models is often not clear, and the models have been developed with little reference to related theory such as in the animal punishment literature (Clutton-Brock and Parker 1995; Clutton-Brock 1998) and Frank's (1998, 2003) recent synthesis of social evolution theory. The basic idea is that if punishment is sufficiently frequent and harsh, it can successfully maintain cooperative behavior. However, this solution forces us to consider the motivation of the punisher. Since a behavior that promotes a public good such as cooperation is in itself a second-order public good and is not expected to be without cost to the actor, punishment is open for exploitation by second-order free-riding individuals who cooperate but who fail to punish defectors (Oliver 1980). Punishment of second-order free riders can be invoked, but this opens up the possibility of third- and higher-order free riding (Ostrom 1990). Failure to maintain participation in a high-level public-goods game unravels participation in the lower levels. At first glance, punishment seems not to be a helpful addition to the problem of cooperation because all that is achieved is the replacement of one public-goods dilemma for another.

[*] Corresponding author; e-mail: andy.gardner@ed.ac.uk.

[†] E-mail: stu.west@ed.ac.uk.

Am. Nat. 2004. Vol. 164, pp. 753–764. © 2004 by The University of Chicago. 0003-0147/2004/16406-40262$15.00. All rights reserved.

Table 1: A simple classification of some mechanisms that promote the evolution of cooperative behaviors

Selection pressure	Fundamental concept	Costs	Benefits
Kin selection	Relatedness between social partners	Cost for actor	Benefit for recipient
Reciprocal altruism	Repression of competition	Cost for actor	Future benefit for actor
Policing	Repression of competition	Cost for actor	Benefits for group
Punishment	Repression of competition	Cost for actor and recipient	Indirect benefit through increased cooperation

However, it is generally true that punishment is cheap relative to the cost of cooperation. Consequently, it has been argued that any mechanism invoked to explain participation in public-goods games will more easily favor punishing (and hence also cooperation) than it would cooperation alone (Sober and Wilson 1998).

A Darwinian account of the evolution of cooperation through punishment requires that the punisher directly or indirectly receives a net benefit through punishing. Although costly punishment can ultimately enhance the direct fitness of the punisher if interactions tend to be extended or repeated with the same social partner (Frank 2003; e.g., sanctioning in plant-rhizobium mutualisms: Denison 2000; West et al. 2002*b*, 2002*c*; Kiers et al. 2003), animals including humans punish even when there is no mechanism ensuring repeat encounters (Fehr and Gächter 2002). Genealogical relationship between social partners is often considered low or absent, and so kin selection is given little attention in the existing literature. The favored Darwinian mechanisms that have received the most attention are group selection (Gintis 2000) and cultural group selection (Heinrich and Boyd 2001). A recent simulation study (Boyd et al. 2003) has suggested that since the incidence of defection declines as punishment becomes more frequent, the costs of punishment decline as it becomes common, so that even modest group selection may plausibly maintain punishment in humans.

In this article, we show that the evolution of punishment and cooperation may be investigated using the powerful direct fitness maximization techniques of Taylor and Frank (1996) and Frank (1998). This allows us to clarify the mechanisms at work and link previous theory to Frank's (1998, 2003) general framework. In particular, we link kin selection, group selection, and cultural group selection in terms of a generalized view of relatedness. We then reveal that it is not the relatedness between social partners per se that facilitates the evolution of punishing behavior. What is crucial is that there is a positive correlation between the punishment strategy played and cooperation received by an individual. Although such an association could arise from viscous population structure and interactions between kin, it may arise for other reasons. In particular, we demonstrate that even in the absence of relatedness it is possible for such an association, due to facultative adjustment of cooperative behavior, to maintain punishment through selection acting at the level of the individual, rendering group selection and elaborate cultural practices unnecessary. More generally, the fact that a positive correlation between the behaviors of social partners is the fundamental factor favoring cooperation has been obscured by a focus on how this correlation can be produced by kinship, through the interactions of close relatives (Hamilton 1975; Frank 1998). Our results provide a clear example of how such positive correlations can arise without kin association.

Models and Analyses

Basic Model

We now present a simple model describing the coevolution of cooperation and punishment. This is intended to elucidate the general selection pressures involved—it is the simplest model that captures the essentials of the problem. We discuss our model in terms of humans because this is where much of the recent theoretical literature has been focused. However, the implications are general and could be applied to a variety of organisms. A role for punishment in the evolution of cooperation has been suggested in a variety of animals, including insects, birds, primates, and other mammals (Clutton-Brock and Parker 1995). We give some specific examples in the discussion when considering how our model may be tested empirically.

For simplicity, we suppose that individuals interact in pairs, with one (random) member of the pair being denoted player 1 and the other player 2. Player 1 may choose to cooperate (e.g., sharing food), in which case she loses fitness c and player 2 gains fitness b, or to defect (e.g., refusing to share food), such that neither player loses nor gains fitness from the interaction. Player 2 may respond to defection in two ways: either she punishes (e.g., by physically injuring player 1) at a cost a to herself in order to reduce player 1's fitness by d, or else she forgives (e.g., does nothing) in which case neither player gains nor loses fitness. The expected direct fitness of a focal individual might then be written as

$$w = \alpha - cx + bX - (1 - X)ya - (1 - x)Yd, \tag{1}$$

where the constant α is baseline fitness, x is the frequency with which that individual cooperates, X is the mean frequency of cooperation among her social partners, y is the frequency with which the individual punishes, given that her partner defects, and Y is the mean punishment strategy played by her social partners, that is, the probability that the focal individual is punished given that she defects. We assume that all competition is global. An important point is that punishment acts to directly reduce both the fitness of the actor and the fitness of her social group. Punishment is therefore fundamentally different from the policing models of Frank (1995, 1996, 2003) because policing directly reduces actor fitness but increases group fitness.

Coevolution of Cooperation and Punishment

We will consider the simultaneous evolutionary optimization of cooperation and punishment analogous to the evolution of policing analysis of Frank (1995), using the direct (neighbor-modulated) fitness maximization method of Taylor and Frank (1996) and Frank (1998). A small increase in a behavior is favored by selection if the derivative of fitness with respect to that behavior (termed "marginal fitness") is >0 and disfavored when this derivative is <0. Differentiating the focal individual's fitness function (eq. [1]) with respect to her cooperating (x) and punishing (y) strategies obtains

$$\frac{\mathrm{d}w}{\mathrm{d}x} = -c + Yd + \frac{\mathrm{d}X}{\mathrm{d}x}(b + ya) - \frac{\mathrm{d}y}{\mathrm{d}x}(1 - X)a - \frac{\mathrm{d}Y}{\mathrm{d}x}(1 - x)d, \tag{2a}$$

$$\frac{\mathrm{d}w}{\mathrm{d}y} = -(1 - X)a - \frac{\mathrm{d}Y}{\mathrm{d}y}(1 - x)d + \frac{\mathrm{d}x}{\mathrm{d}y}(Yd - c) + \frac{\mathrm{d}X}{\mathrm{d}y}(b + ya). \tag{2b}$$

The terms $\mathrm{d}X/\mathrm{d}x$ and $\mathrm{d}Y/\mathrm{d}y$ are the coefficients of relatedness, with respect to cooperation and punishment, respectively, between the focal individual and her social partners (Taylor and Frank 1996; Frank 1998). Technically, the derivative is of the conditional expectation of the social partner's strategy, given the strategy played by the focal individual, with respect to the latter. The other derivative terms are $\mathrm{d}y/\mathrm{d}x$ and $\mathrm{d}x/\mathrm{d}y$, which are the regression of an individual's punishing strategy on its own cooperation strategy, and vice versa, and $\mathrm{d}Y/\mathrm{d}x$ and $\mathrm{d}X/\mathrm{d}y$, which are the regressions of a partner's punishing strategy on its own cooperation strategy and a partner's cooperation strategy on its own punishment strategy, respectively.

Let us consider first the origin of cooperation and punishment in a population that is otherwise fixed for defection ($\bar{x} \rightarrow 0$) and forgiveness ($\bar{y} \rightarrow 0$). In such circumstances the trait-on-trait regressions are always nonnegative, which is important for interpretation of the analytical results that follows. To see why, consider the regression of cooperation received on cooperation strategy played: $\mathrm{d}X/\mathrm{d}x = (X - \bar{x})/(\mathrm{x} - \bar{x}) \approx X/x$. Since cooperation strategies are nonnegative, the numerator (X) is nonnegative, and since the variant by definition plays a different cooperation strategy from the wild type (which plays zero cooperation), the denominator (x) is positive. Hence, $\mathrm{d}X/\mathrm{d}x \geq 0$. The same argument can be used to show that this is true for the other trait-on-trait regressions. Assuming only minor variants ($x \approx X \approx \bar{x}$, $y \approx Y \approx \bar{y}$; Taylor and Frank 1996; Frank 1998) and making the substitutions $\bar{x} \rightarrow 0$ and $\bar{y} \rightarrow 0$, the marginal fitness with respect to cooperation (eq. [2a]) reduces to

$$\frac{\mathrm{d}w}{\mathrm{d}x} = -c + \frac{\mathrm{d}X}{\mathrm{d}x}b - \frac{\mathrm{d}y}{\mathrm{d}x}a - \frac{\mathrm{d}Y}{\mathrm{d}x}d. \tag{3}$$

This shows there is a direct cost (c) and a kin-selected benefit ($\mathrm{d}X/\mathrm{d}x \times b$) of cooperation, plus costs relating to the associated increase in costly punishing ($\mathrm{d}y/\mathrm{d}x \times a$) and also in being punished ($\mathrm{d}Y/\mathrm{d}x \times d$); see figure 1*A*. Cooperation is maintained even in the absence of punishment when Hamilton's (1964) rule $\mathrm{d}X/\mathrm{d}x \times b > c$ holds, so we will consider the more interesting situation where it does not, such that equation (3) is always negative.

Similarly, the marginal fitness with respect to punishment (eq. [2b]) is

$$\frac{\mathrm{d}w}{\mathrm{d}y} = -a - \frac{\mathrm{d}Y}{\mathrm{d}y}d - \frac{\mathrm{d}x}{\mathrm{d}y}c + \frac{\mathrm{d}X}{\mathrm{d}y}b. \tag{4}$$

Again, this is easily understood. Punishing incurs a direct cost (a) and indirect costs ($\mathrm{d}Y/\mathrm{d}y \times d$ from being punished by related individuals and $\mathrm{d}x/\mathrm{d}y \times c$ from the correlated commitment to cooperation). The benefit $\mathrm{d}X/\mathrm{d}y \times b$ is gained through the association between the punishment strategy played and the cooperation received (see fig. 1*B*). Only when this is sufficiently large may a rare variant with some small frequency of punishing behavior be able to invade. In other words, a positive association between the punishment strategy played and the cooperation received by a focal individual is a necessary but not sufficient condition for the evolutionary origin of punishment.

Result 1. A positive association between punishment strategy played and cooperation received is crucial for the evolutionary origin of punishing behavior.

We will now investigate the evolutionary maintenance of cooperation and punishment by considering $\bar{x} \rightarrow 1$ and

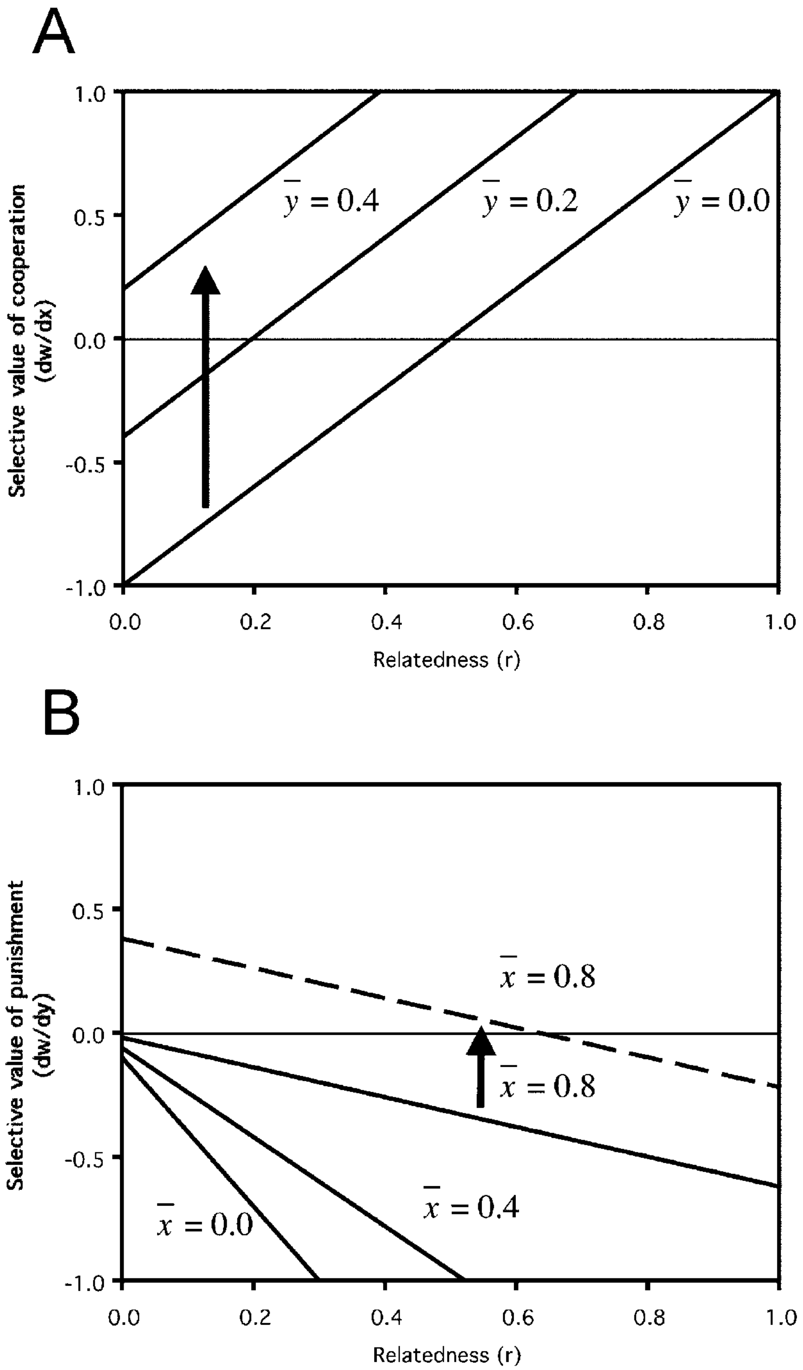

Figure 1: *A*, Selective value of cooperation (dw/dx) as a function of relatedness and the resident punishing strategy ($\bar{y}$) when there is no association between traits ($dy/dx = dY/dx = 0$); $dw/dx > 0$ indicates that enhanced cooperation is favored, and $dw/dx < 0$ indicates that it is disfavored. Increasing relatedness (r) enhances selection for cooperation; in the absence of punishment, cooperation is favored when $rb > c$. Increasing punishment also favors cooperation, so cooperation may be favored even when relatedness is 0, if $\bar{y} > c/d$. *B*, Selective value of punishment (dw/dy) as a function of relatedness and the resident cooperation strategy ($\bar{x}$); $dw/dy > 0$ indicates enhanced punishment is favored, and $dw/dy < 0$ indicates that it is disfavored. Assuming no association between traits ($dx/dy = dX/dy = 0$), we see that punishment is always disfavored, that increased relatedness enhances the selective disadvantage of punishment, and that increased cooperation reduces the selective disadvantage of punishment. Punishment may be favorable if there is a positive association between the punishment strategy played and the cooperation received by an individual ($dX/dy > 0$); the broken line indicates $dX/dy = 0.2$. For *A* and *B*, we assume $a = 0.1$, $b = 2$, $c = 1$, and $d = 3$.

$\bar{y}\rightarrow 1$. Again, the trait-on-trait regressions will all be nonnegative: for example, $\mathrm{d}X/\mathrm{d}x = (X-\bar{x})/(x-\bar{x}) \approx (X-1)/(x-1)$. Cooperation received ($X$) cannot be >1, so the numerator ($X-1$) is ≤ 0. Since the cooperation variant does not play the wild-type strategy (always cooperate) and cannot play a more cooperative strategy than that, the denominator ($x-1$) is always negative. Hence, $\mathrm{d}X/\mathrm{d}x \geq 0$. Making the substitutions $\bar{x}\rightarrow 1$ and $\bar{y}\rightarrow 1$, the marginal fitness with respect to cooperation (eq. [2a]) is now given by

$$\frac{\mathrm{d}w}{\mathrm{d}x} = -c + d + \frac{\mathrm{d}X}{\mathrm{d}x}(b+a). \tag{5}$$

Here cooperation carries a direct cost (c) and a benefit (d, due to avoiding punishment) when punishment of defectors is assured. It also gives kin-selected benefits ($\mathrm{d}X/\mathrm{d}x \times b$ and $\mathrm{d}X/\mathrm{d}x \times a$) due to the correlated cooperation received from social partners and the fitness saved from not having to punish defectors. Punishment cannot be an effective deterrent when the fitness of a punished defector is greater than that of a cooperator, so that we will restrict attention to the situation $d > c$. Here, the marginal fitness will always be positive, and so selection will act to maintain cooperation. The marginal fitness with respect to punishment (eq. [2b]) is

$$\begin{aligned}\frac{\mathrm{d}w}{\mathrm{d}y} = &-(1-\bar{x})a - \frac{\mathrm{d}Y}{\mathrm{d}y}(1-\bar{x})d \\ &+ \frac{\mathrm{d}x}{\mathrm{d}y}(d-c) + \frac{\mathrm{d}X}{\mathrm{d}y}(b+a).\end{aligned} \tag{6}$$

The costs of punishment include the direct cost ($[1-\bar{x}] \times a$) and the kin-selected cost ($[1-\bar{x}] \times \mathrm{d}Y/\mathrm{d}y \times d$) plus the cost incurred by the associated cooperation ($\mathrm{d}x/\mathrm{d}y \times c$). The benefits of punishment are due to the correlated decrease in one's own defection and hence the frequency with which the focal individual is punished ($\mathrm{d}x/\mathrm{d}y \times d$) and also the correlated increase in cooperation received from social partners ($\mathrm{d}X/\mathrm{d}y \times b$) and, conversely, the fitness saved by not having to punish partners ($\mathrm{d}X/\mathrm{d}y \times a$). If $\mathrm{d}x/\mathrm{d}y = \mathrm{d}X/\mathrm{d}y = 0$ so that there is no correlation between the punishment and cooperation played by an individual, nor between the punishment played and cooperation received, then the marginal fitness with respect to punishment is small but negative, and hence full punishment is not stable. It is interesting to note that relatedness $\mathrm{d}Y/\mathrm{d}y$ works to undermine the stability of punishment; as an individual's punishment strategy is increased, so too is the punishment received from social partners. If the between-trait associations are positive and of sufficient magnitude, then full punishment can be evolutionarily stable. Otherwise, selection will act to reduce punishment in the population.

Result 2. A positive association between punishment strategy played and cooperation received is crucial for the evolutionary maintenance of punishing behavior.

We now check to see whether punishment is easier to maintain than it is to initially invade an otherwise forgiving population, by evaluating $\mathrm{d}w/\mathrm{d}y|_{\bar{x},\bar{y}=1} - \mathrm{d}w/\mathrm{d}y|_{\bar{x},\bar{y}=0}$, that is, subtracting the right-hand side (RHS) of equation (4) from the RHS of equation (6) to obtain

$$d\left(\frac{\mathrm{d}Y}{\mathrm{d}y} + \frac{\mathrm{d}x}{\mathrm{d}y}\right) + a\left(1 + \frac{\mathrm{d}X}{\mathrm{d}y}\right), \tag{7}$$

which is positive, so that RHS equation (4) is less than RHS equation (6), and hence the condition for increased punishment to be favored ($\mathrm{d}w/\mathrm{d}y > 0$) is more easily satisfied in a population of cooperators and punishers than in a population of defectors and forgivers. Similarly, the RHS of equation (3) is always negative under the relevant circumstances (i.e., when $\mathrm{d}X/\mathrm{d}x \times b < c$), and the RHS of equation (5) is always positive, so that the condition for enhanced cooperation to be favored ($\mathrm{d}w/\mathrm{d}x > 0$) is also more easily satisfied in punishing populations than in populations rife with defection and forgiveness.

Result 3. Punishing behavior is more easily maintained than it is originally evolved. Note that this assumes that relatedness and the between-trait regressions are constants. A fully dynamic analysis relaxing this assumption would require that we specify a more detailed (and hence less general) model and so is not pursued here because we aim only to abstract and elucidate the selection pressures involved in the evolution of punishment and cooperation.

Example: Cooperation as a Facultative Response to Punishment

The Model. We have found that relatedness between social partners is not crucial for costly punishment to be favored (indeed, increasing relatedness disfavors punishment) and that it is another association, the regression of the cooperation received on the punishment strategy played, that provides the benefit of punishment. To illustrate these findings, we examine the evolution of punishment when there is no relatedness between individuals ($\mathrm{d}Y/\mathrm{d}y = 0$) and when cooperation is facultatively adjusted to one's punishment environment (which we will see can give $\mathrm{d}X/\mathrm{d}y > 0$).

We assume that individuals are randomly organized into social groups of size N, such that relatedness between group members is 0. In each social encounter, individuals pair with a random member from their group, with one of the partners playing the role of player 1 and the other being player 2. In contrast with the previous model, we consider the cooperation strategy of player 1 to be facultative and hence a function of her punishment environment. Assuming no partner recognition and therefore no adjustment of cooperation to her current partner's punishment strategy, the cooperation strategy played by the focal individual (in half of her social interactions) is expressed as a function of the average punishment strategy played by all of her social partners: $x = f(\bar{y})$. Since each of her social partners experiences a punishing environment that includes the focal individual (and hence average punishment strategy among their social partners is $\bar{y} + [y - \bar{y}]/[N - 1]$), they will play cooperation strategy $X = f(\bar{y} + (y - \bar{y})/(N - 1))$.

If individuals cooperate optimally, we expect the function $f(Y)$ to be such that it maximizes the fitness of player 1 when player 2 plays punishment strategy Y. It is easy to show that this optimum is given by

$$f^*(Y) = \begin{cases} 0 & c > Yd \\ \quad \mathit{if} & \\ 1 & c < Yd \end{cases}, \tag{8}$$

such that defection is favored when the cost of cooperation outweighs the threat of punishment ($c > Yd$), and cooperation is favored when the cost of cooperation is outweighed by the threat of punishment ($c < Yd$). This step function is both mathematically inconvenient and biologically unreasonable, so we will use the model of McNamara et al. (1997; see also Kokko 2003) to describe nearly optimized cooperation as

$$f(Y) = \frac{1}{1 + \exp(-\Delta/\varepsilon)} = \frac{1}{1 + \exp[-(Yd - c)/\varepsilon]}, \tag{9}$$

where ε is the degree of behavioral error and $\Delta = \mathrm{d}w/\mathrm{d}x = Yd - c$ ensures that the frequency of nonoptimal behavior declines as its impact on fitness becomes more important. The facultative cooperation function (eq. [9]) approaches the step function (eq. [8]) for vanishing behavioral error ($\varepsilon \rightarrow 0$), and for larger error ($\varepsilon > 0$), it takes a continuous sigmoidal form which flattens out to a constant 1/2 as the error tends to infinity (fig. 2). For mathematical convenience, we will assume vanishing (but nonzero) behavioral error ($\varepsilon \rightarrow 0$).

Altering fitness function (eq. [1]) for this example model, we have the fitness of an individual who plays punishment strategy y, in a population with mean punishment strategy $\bar{y}$, given by

$$\begin{aligned} w = \alpha - cf(\bar{y}) + bf\left(\bar{y} + \frac{(y - \bar{y})}{(N - 1)}\right) \\ - a\left[1 - f\left(\bar{y} + \frac{(y - \bar{y})}{(N - 1)}\right)\right]y \\ - d(1 - f(\bar{y}))\bar{y}. \end{aligned} \tag{10}$$

The mean fitness of the population is

$$\begin{aligned} \bar{w} = \alpha - cf(\bar{y}) + bf(\bar{y}) \\ - a(1 - f(\bar{y}))\bar{y} - d(1 - f(\bar{y}))\bar{y}, \end{aligned} \tag{11}$$

so we expect a rare variant playing punishment strategy y to increase in frequency in a population with mean punishment strategy $\bar{y}$ when the fitness differential $\Delta w = w - \bar{w}$ is positive, that is, when

$$\begin{aligned} \Delta w = b\left[f\left(\bar{y} + \frac{(y - \bar{y})}{(N - 1)}\right) - f(\bar{y})\right] \\ - a\left[1 - f\left(\bar{y} + \frac{(y - \bar{y})}{(N - 1)}\right)y - (1 - f(\bar{y})\bar{y}\right] > 0. \end{aligned} \tag{12}$$

Origin of Punishment. We first consider the evolutionary stability (Maynard Smith and Price 1973) of forgiveness, by determining under what circumstances no variant with punishment strategy $y > 0$ can invade a population with mean punishment strategy $\bar{y} \rightarrow 0$. Substituting the cooperation function (eq. [9]) into the fitness differential (eq. [12]) obtains

$$\begin{aligned} \Delta w = b\left[\frac{1}{1 + \exp(\{c - [y/(N - 1)]d\}/\varepsilon)} - \frac{1}{1 + \exp(c/\varepsilon)}\right] \\ - a\left[1 - \frac{1}{1 + \exp(\{c - [y/(N - 1)]d\}/\varepsilon)}\right]y. \end{aligned} \tag{13}$$

Recalling that the behavioral error is vanishingly small ($\varepsilon \rightarrow 0$), we find that when the threat of punishment posed to social partners of the punishing variant is less than the cost of cooperation ($[yd]/[N - 1] < c$), then equation (13) reduces to $-y\ a$, which is negative, and hence the rare variant cannot invade. This is because defection is the rule in the social groups of both the wild type and the variant, giving population mean fitness $\bar{w} \approx \alpha$ and rare variant fitness $w \approx \alpha - ya$. When the threat of punishment is greater than the cost of cooperation ($[yd]/[N - 1] > c$), then equation (13) reduces to b, which is positive, and hence the rare variant can invade. Here, the rare punisher has managed to push her social group over the punishment threshold such that cooperation is now the optimal strat-

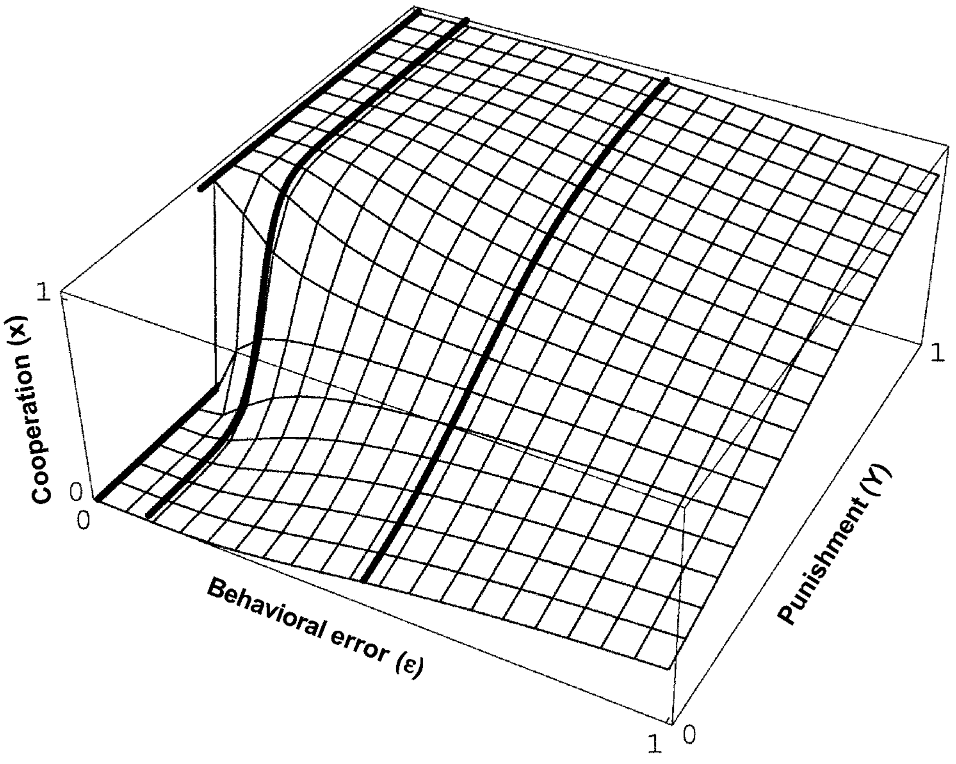

Figure 2: Frequency with which an individual cooperates (x) as a function of the punishment strategy of its social partners (Y) and the degree of behavioral error (ε), according to the example facultative model. Values are obtained numerically, assuming $c = 1$ and $d = 3$. The bold lines indicate $\varepsilon = 0$, 0.1, and 0.5.

egy. The average social group is fully defecting, so $\bar{w} \approx \alpha$, but the rare variant is now a recipient of cooperative behavior and only rarely encounters a defector requiring punishment, so that her fitness is $w \approx \alpha + b$. Note that although the variant receives cooperation, she maximizes her fitness by always defecting (since her unrelated social partners are all forgivers) and hence pays no cost of cooperation. If no y satisfies the above invasion condition, then forgiveness is an evolutionarily stable strategy (ESS; Maynard Smith and Price 1973). This is assured when $(N - 1)c > d$, so that not even a fully punishing variant ($y = 1$) can invade. Evolutionary stability of forgiveness is therefore assured unless

$$d > (N - 1)c. \tag{14}$$

Result 4. In the above model, punishment is unlikely to invade forgiveness unless the population is structured into very small groups.

Maintenance of Punishment. To determine whether punishment is an ESS, we let the wild type adopt the strategy of full punishment ($\bar{y} \rightarrow 1$) and consider the success of rare variants playing $y < 1$. Substituting the facultative cooperation function (eq. [9]) into the fitness differential (eq. [12]) obtains

$$\Delta w = b\left\{\frac{1}{1 + \exp(\{c - [1 - (1 - y)/(N - 1)]d\}/\varepsilon)} - \frac{1}{1 + \exp[(c - d)/\varepsilon]}\right\} + a\left\{1 - \frac{1}{1 + \exp[(c - d)/\varepsilon]} - \left[1 - \frac{1}{1 + \exp(\{c - [1 - (1 - y)/(N - 1)]d\}/\varepsilon)}\right]y\right\}. \tag{15}$$

First consider "ineffective punishment" ($c > d$). When behavioral error is vanishing ($\varepsilon \rightarrow 0$), the fitness differential (eq. [15]) reduces to $a(1 - y)$, which is positive, and hence the more forgiving variant will always invade. This is because even when defection is always met with punishment, the defector has greater fitness than the cooperator, so that in all social groups defection is rife. The resident strategy incurs the cost of full punishment, and so the mean fitness of the population is $\bar{w} \approx \alpha - a$, whereas the more forgiving variant avoids this at least part of the time, giving fitness $w \approx \alpha - ya$. Now consider "effective punishment" ($d > c$),

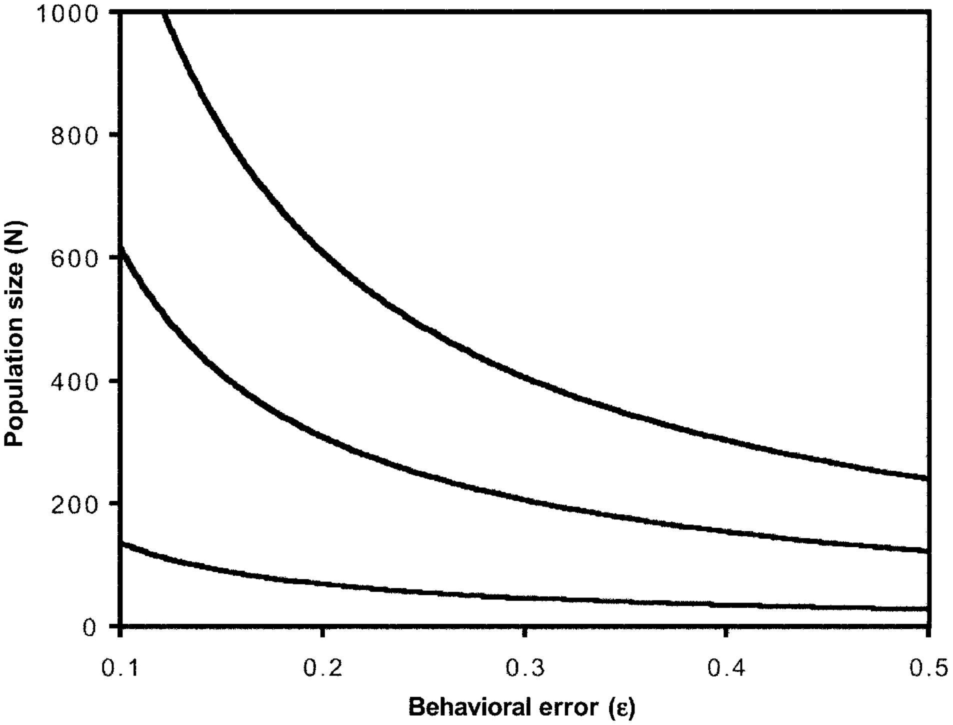

Figure 3: Maximum group size (N) permitting the evolutionary stability of punishment ($\bar{y} = 1$) as a function of behavioral error (ε) and the cost of punishing (a), according to the example facultative model, assuming $b = 2$, $c = 1$, and $d = 3$. Upper line, $a = 0.01$; middle line, $a = 0.10$; bottom line, $a = 0.50$.

such that punished defectors receive lower fitness than cooperators. The resident now enjoys the benefits of cooperation and only infrequently encounters erroneous defection requiring punishment. If the rare variant forgives to such a degree that her social partners optimize by defection; that is, when $c - [1 - (1 - y)/(N - 1)]d > 0$, the fitness differential (eq. [15]) reduces to $-(b + ya)$ since she loses the benefits of cooperation and punishes a proportion y of her social partners. This is negative, and so the rare variant cannot invade. If the variant's forgiveness is not sufficient to warrant a switch to defection among her social partners, equation (15) becomes $-(b + ya) \exp\{c - [1 - (1 - y)/(N - 1)]d\}$, which is vanishingly small but nevertheless negative, and hence the rare variant cannot invade. This is true because with vanishing behavioral error ($\varepsilon \rightarrow 0$) the frequency of defection in the fully punishing group is a vanishing fraction of the frequency of defection in the more forgiving group, so that the fitness saved from not punishing so frequently does not outweigh the fitness lost through the reduction of received cooperation. Relaxation of the infinitesimal error assumption (fig. 3) shows that this result is robust, even for large social groups. The variant can therefore only invade an otherwise fully punishing population when punishment is ineffective, so that punishment is an ESS when

$$d > c. \qquad (16)$$

Result 5. In the above model, punishment is maintained by selection once it has become common if the cost of cooperation (c) is less than the cost of being punished (d).

Discussion

Punishment and Cooperation

We have shown that full punishment can be an evolutionarily stable strategy only if there is a positive association between the punishment played and the cooperation received by an individual. This could arise if populations are viscous so that social partners tend to be genealogical relatives, but other mechanisms are possible, for example, when individuals facultatively adjust their level of cooperation in response to the local threat of punishment. We have also provided analytical support for the suggestion of Boyd et al. (2003) that the cost of punishment declines as it becomes common in the population and hence punishing behavior might be maintained more easily than it is initially evolved.

These results suggest three general implications. First, it can be easier for some cooperation to evolve by another mechanism (e.g., altruism between relatives) and then punishment evolve to favor and maintain higher levels of cooperation. An analogous conclusion has been made for some other mechanisms that do not rely on interactions between relatives, such as group augmentation (Kokko et al. 2001; Griffin and West 2002). Second, within the specific context of explaining human cooperation, punishment could have evolved at a time when social structure was more conducive to punishment (small groups of interacting individuals). Once common, punishment could be retained even when interactions began to occur within much larger groups of humans. Third, the opposite frequency dependence is true for systems based on rewarding cooperation rather than punishing defection—the cost of rewarding escalates as more individuals cooperate, whereas we have shown the cost of punishing decreases as more individuals cooperate. This might go some way to explaining why punishment as opposed to rewarding is prevalent in nature (e.g., Clutton-Brock and Parker 1995).

How can our model be tested? Our major result is that costly punishment can be favored if there is a positive association between the punishment played and the cooperation received by an individual (results 1 and 2). This could be hard to test directly, especially experimentally, because of limitations on how an individual's level of punishment could be manipulated. However, some of the fundamental assumptions and predictions of our model that underly this result could be tested more easily. In particular, are lower levels of cooperation more likely to lead to punishment, as appears to occur in superb fairy wrens (Mulder and Langmore 1993), naked mole rats (Reeve 1992), and *Polistes* wasps (Reeve and Gamboa 1987)? Second, are individuals more likely to cooperate when they are punished, as may occur in *Polistes* wasps (Reeve and Gamboa 1987)? Third, do individuals try to signal that they cooperate more than they actually do, as occurs in white-winged choughs (Boland et al. 1997)? Fourth, do systems in which social partners are more related tend to display less punishment, analogous with Frank's (1995, 2003) result that investment into policing correlates negatively with relatedness?

Relatedness and Kin Selection

This analysis has made use of the understanding that the coefficient of relatedness appropriate to the direct fitness formulation of Hamilton's rule is a regression measure describing the association between actor and social partner phenotypes (reviewed by Seger 1981; Michod 1982; Grafen 1985; Queller 1985; 1992; Frank 1998). Such associations are generally due to genealogical closeness and hence genetic similarity, so that the maximization of neighbor-modulated or inclusive fitness is popularly referred to as "kin selection" (Maynard Smith 1964). Group selection can be responsible for the evolution of an altruistic trait only insofar as the benefit to the group is large enough, the cost to the individual is low enough, and there is substantial between-group as opposed to within-group variation in trait values. Since the proportion of the total variance that is attributable to between-group differences is the coefficient of relatedness appropriate for whole-group traits, Hamilton's rule can be used to predict when group selection will favor the trait (i.e., when relatedness × benefit > cost). Thus, kin selection and group selection are mathematically equivalent ways of conceptualizing the same evolutionary process, a point that previously has been analyzed in much detail (Price 1972; Hamilton 1975; Wade 1985; Frank 1986, 1998; Queller 1992; Reeve and Keller 1999). Consequently, it is puzzling that kin selection has been largely ignored in the human altruistic punishment literature on the grounds that relatedness is too low, while group selection has often been regarded as important (e.g., Gintis 2000). Furthermore, because relatedness is a regression of recipient phenotype on actor phenotype, it transcends genetics and applies even when the cause of phenotypic similarity is simply imitation, for example, as in the cultural group selection proposed by Heinrich and Boyd (2001). In this sense, "kin selection" is something of a misnomer because it draws attention to only one cause of the statistical association that is relatedness, as Hamilton (1975) realized.

As this analysis has shown, positive relatedness is not really the key ingredient for the evolutionary success of punishment. Punishing behavior is costly to the individual and protects the social group from the breakdown of cooperation, and hence it has been described as a form of altruism (Sober and Wilson 1998). It might then be expected that where it is successful, altruistic punishment is being maintained by kin selection. However, punishment is quite a different form of public good from cooperation—it is directly disadvantageous at the group level because it reduces the fitness of the focal individual and her social partners. The benefit it brings is indirect because it merely creates a coercive social environment in which cooperation is favored. It therefore differs from Frank's (1995, 1996, 2003) recent models of competition-repression in which investment into policing behavior translates directly into enhanced group fitness. In our model, punishment is only of selective value when there is a sufficiently strong correlation between punishment strategy played and cooperation received ($\mathrm{d}X/\mathrm{d}y$; fig. 1*B*).

This highlights a fundamental nonequivalence of first- and higher-order public goods.

A positive correlation between punishment played and cooperation received might arise in a viscous population where genealogical kin tend to associate with each other, so that the social partners of punishers are also punishers ($\mathrm{d}Y/\mathrm{d}y > 0$) and therefore punishers are expected to be coerced into cooperating more than forgivers ($\mathrm{d}x/\mathrm{d}y > 0$). This association combines with relatedness to ensure that an increase in punishing behavior is associated with an increase in the amount of cooperation received ($\mathrm{d}X/\mathrm{d}y > 0$). The pressure for enhanced punishment is therefore not strictly kin selection but rather something more akin to "niche construction" (Odling-Smee et al. 1996), in the sense that the behavior modifies the social environment in such a way as to alter the selective pressures acting upon other traits. It is worth noting that localized competition in viscous populations adds extra complexity to models of kin selection (see Taylor 1992*a*, 1992*b*; Wilson et al. 1992; Queller 1994; Frank 1998; Griffin and West 2002; West et al. 2002*a*; Gardner and West 2004 for extensive discussion of its impact on the evolution of social behaviors). In our analysis, we have assumed that all competition occurs at the level of the whole population, and we leave local competition as an open problem for the future.

We may easily demonstrate that relatedness is not necessary for the evolution of costly punishment by considering mechanisms that generate positive associations between the punishment played and the cooperation received despite zero relatedness, for example, the facultative model of cooperation introduced above. We discovered that in the absence of relatedness, partner recognition, reputation, and any mechanism whereby an individual may bias her interactions or tailor her behavior in response to her immediate social partner, punishment might be maintained by selection acting directly at the level of the individual. This is because when punishment is already frequent, the fitness saved by forgiving is minimal and may be overwhelmed by the concomitant decline in the amount of cooperation received because of the decrease in selection for cooperation among social partners. This example model is intended for illustration only and is designed to demonstrate how a net benefit for punishment might be achieved even when individuals do not interact with relatives. More complicated scenarios are therefore possible, and of particular interest is the effect of enhanced behavioral error (increasing ε). Numerical analysis of the example model reveals that increasing the frequency of maladaptive behavior reduces the likelihood that individual level selection will be able to maintain altruistic punishment in very large groups (fig. 3), although the results presented above are qualitatively robust so long as behavioral error (ε) and the cost of punishing (a) are small. The degree to which individuals are expected to behave optimally is contentious, but punishment is indeed characterized by its cheapness (Sober and Wilson 1998).

Conclusion

We have given analytical support to the suggestion that the cost of punishment declines as it becomes a common strategy, so that punishment is more easily maintained than it is originally evolved. We showed that it is not relatedness per se that is important in ensuring that punishing behavior enhances fitness but rather that a positive correlation between punishment played and cooperation received by an individual is crucial. We also revealed that facultative adjustment of cooperation can give rise to such a positive association even in the absence of relatedness between social partners. Finally, we demonstrated that the direct benefits accrued when cooperation is facultative may be large enough for selection acting at the individual level alone to maintain punishment among humans, rendering elaborate population dynamics and cultural practices unnecessary. More generally, our results provide a specific example of how positive correlations between the behaviors played by social partners can arise and favor cooperation for reasons other then kinship. Major tasks for the future include clarifying the links between punishment and reproductive skew theory (Johnstone 2000; Clutton-Brock et al. 2001; Langer et al. 2004) and developing more specific models for specific situations or organisms.

Acknowledgments

We thank N. Barton, T. Clutton-Brock, F. Dionisio, S. Frank, A. Kalinka, H. Kokko, S. Nee, J. Pepper, D. Wilson, and W. Zuidema for their comments on the manuscript. Funding was provided by the Biotechnology and Biological Sciences Research Council, the Natural Environment Research Council, and the Royal Society.

Literature Cited

Alexander, R. D. 1979. Darwinism and human affairs. University of Washington Press, Seattle.

———. 1987. The biology of moral systems. Aldine de Gruyter, New York.

Boland, C. R. J., R. Heinsohn, and A. Cockburn. 1997. Deception by helpers in cooperatively breeding white-winged choughs and its experimental manipulation. Behavioral Ecology and Sociobiology 41:251–256.

Boyd, R., and P. J. Richerson. 1992. Punishment allows the evolution of cooperation (or anything else) in sizable groups. Ethology and Sociobiology 13:171–195.

Boyd, R., H. Gintis, S. Bowles, and P. J. Richerson. 2003.

The evolution of altruistic punishment. Proceedings of the National Academy of Sciences of the USA 100:3531–3535.

Buss, L. W. 1987. The evolution of individuality. Princeton University Press, Princeton, N.J.

Clutton-Brock, T. H. 1998. Reproductive skew, concessions and limited control. Trends in Ecology & Evolution 13: 288–292.

Clutton-Brock, T. H., and G. A. Parker. 1995. Punishment in animal societies. Nature 373:209–216.

Clutton-Brock, T. H., P. N. M. Brotherton, A. F. Russell, M. J. O'Riain, D. Gaynor, R. Kansky, A. Griffin, et al. 2001. Cooperation, control, and concession in meerkat groups. Science 291:478–481.

Denison, R. F. 2000. Legume sanctions and the evolution of symbiotic cooperation by rhizobia. American Naturalist 156:567–576.

Fehr, E., and S. Gächter. 2000. Cooperation and punishment in public goods experiments. American Economic Review 90:980–994.

———. 2002. Altruistic punishment in humans. Nature 415:137–140.

Frank, S. A. 1986. Hierarchical selection theory and sex ratios I. General solutions for structured populations. Theoretical Population Biology 29:312–342.

———. 1995. Mutual policing and repression of competition in the evolution of cooperative groups. Nature 377:520–522.

———. 1996. Policing and group cohesion when resources vary. Animal Behaviour 52:1163–1169.

———. 1998. Foundations of social evolution. Princeton University Press, Princeton, N.J.

———. 2003. Repression of competition and the evolution of cooperation. Evolution 57:693–705.

Gardner, A., and S. A. West. 2004. Spite and the scale of competition. Journal of Evolutionary Biology (in press).

Gintis, H. 2000. Strong reciprocity and human sociality. Journal of Theoretical Biology 206:169–179.

Grafen, A. 1985. A geometric view of relatedness. Oxford Surveys in Evolutionary Biology 2:28–89.

Griffin, A. S., and S. A. West. 2002. Kin selection: fact and fiction. Trends in Ecology & Evolution 17:15–21.

Hamilton, W. D. 1963. The evolution of altruistic behavior. American Naturalist 97:354–356.

———. 1964. The genetical evolution of social behavior. I, II. Journal of Theoretical Biology 7:1–52.

———. 1970. Selfish and spiteful behavior in an evolutionary model. Nature 228:1218–1220.

———. 1975. Innate social aptitudes of man: an approach from evolutionary genetics. Pages 133–153 *in* R. Fox, ed. Biosocial anthropology. Malaby, London.

Heinrich, J., and R. Boyd. 2001. Why people punish defectors. Journal of Theoretical Biology 208:79–89.

Hirshleifer, D., and E. Rasmusen. 1989. Cooperation in a repeated prisoner's dilemma with ostracism. Journal of Economic Behavior and Organization 12:87–106.

Johnstone, R. A. 2000. Models of reproductive skew: a review and synthesis. Ethology 106:5–26.

Kiers, E. T., R. A. Rouseau, S. A. West, and R. F. Denison. 2003. Host sanctions and the legume-rhizobium mutualism. Nature 425:78–81.

Kokko, H. 2003. Are reproductive skew models evolutionarily stable? Proceedings of the Royal Society of London B 270:265–270.

Kokko, H., R. A. Johnstone, and T. H. Clutton-Brock. 2001. The evolution of cooperative breeding through group augmentation. Proceedings of the Royal Society of London B 268:187–196.

Langer, P., K. Hogendoorn, and L. Keller. 2004. Tug-of-war over reproduction in a social bee. Nature 428:844–847.

Maynard Smith, J. 1964. Group selection and kin selection. Nature 201:1145–1147.

Maynard Smith, J., and G. R. Price. 1973. The logic of animal conflict. Nature 246:15–18.

Maynard Smith, J., and E. Szathmáry. 1995. The major transitions in evolution. Oxford University Press, Oxford.

McNamara, J. M., J. N. Webb, E. J. Collins, T. Székely, and A. I. Houston. 1997. A general technique for computing evolutionary stable strategies based on errors in decision-making. Journal of Theoretical Biology 189: 211–225.

Michod, R. E. 1982. The theory of kin selection. Annual Review of Ecology and Systematics 13:23–55.

Mulder, R. A., and N. E. Langmore. 1993. Dominant males punish helpers for temporary defection in superb fairy wrens. Animal Behaviour 45:830–833.

Nee, S. 1989. Does Hamilton's rule describe the evolution of reciprocal altruism? Journal of Theoretical Biology 141:81–91.

Odling-Smee, F. J., K. N. Laland, and M. W. Feldman. 1996. Niche construction. American Naturalist 147:641–648.

Oliver, P. 1980. Rewards and punishments as selective incentives for collective action: theoretical investigations. American Journal of Sociology 85:1356–1375.

Ostrom, E. 1990. Governing the commons. Cambridge University Press, New York.

Price, G. R. 1972. Extension of covariance selection mathematics. Annals of Human Genetics 35:485–490.

Queller, D. C. 1985. Kinship, reciprocity, and synergism in the evolution of social behavior. Nature 318:366–367.

———. 1992. Quantitative genetics, inclusive fitness, and group selection. American Naturalist 139:540–558.

———. 1994. Relatedness in viscous populations. Evolutionary Ecology 8:70–73.

Ratnieks, F. L. W. 1988. Reproductive harmony via mutual policing by workers in eusocial Hymenoptera. American Naturalist 132:217–236.

Reeve, H. K. 1992. Queen activation of lazy workers in colonies of the eusocial naked mole-rat. Nature 358: 147–149.

Reeve, H. K., and J. Gamboa. 1987. Queen regulation of worker foraging in paper wasps: a social feedback-control system (*Polistes fuscatus*, *Hymenoptera*, *Vespidae*). Behaviour 102:147–167.

Reeve, H. K., and L. Keller. 1999. Levels of selection: burying the units-of-selection debate and unearthing the crucial new issues. Pages 3–14 *in* L. Keller, ed. Levels of selection in evolution. Princeton University Press, Princeton, N.J.

Seger, J. 1981. Kinship and covariance. Journal of Theoretical Biology 91:191–213.

Sell, J., and R. K. Wilson. 1999. The maintenance of cooperation: expectations of future interaction and the trigger of group punishment. Social Forces 77:1551–1570.

Sigmund, K., C. Hauert, and M. A. Nowak. 2001. Reward and punishment. Proceedings of the National Academy of Sciences of the USA 98:10757–10762.

Sober, E., and D. S. Wilson. 1998. Unto others: the evolution and psychology of unselfish behavior. Harvard University Press, Cambridge, Mass.

Taylor, P. D. 1992*a*. Altruism in viscous populations: an inclusive fitness approach. Evolutionary Ecology 6:352–356.

———. 1992*b*. Inclusive fitness in a heterogeneous environment. Proceedings of the Royal Society of London B 249:299–302.

Taylor, P. D., and S. A. Frank. 1996. How to make a kin selection model. Journal of Theoretical Biology 180:27–37.

Trivers, R. L. 1971. The evolution of reciprocal altruism. Quarterly Review of Biology 46:35–57.

Wade, M. J. 1985. Soft selection, hard selection, kin selection, and group selection. American Naturalist 125: 61–73.

West, S. A., I. Pen, and A. S. Griffin. 2002*a*. Cooperation and competition between relatives. Science 296:72–75.

West, S. A., E. T. Kiers, I. Pen, and R. F. Denison. 2002*b*. Sanctions and mutualism stability: when should less beneficial mutualists be tolerated? Journal of Evolutionary Biology 15:830–837.

West, S. A., E. T. Kiers, E. L. Simms, and R. F. Denison. 2002*c*. Sanctions and mutualism stability: why do rhizobia fix nitrogen? Proceedings of the Royal Society of London B 269:685–694.

Wilson, D. S., G. B. Pollock, and L. A. Dugatkin. 1992. Can altruism evolve in purely viscous populations? Evolutionary Ecology 6:331–341.

Woodcock, S., and J. Heath. 2002. The robustness of altruism as an evolutionary strategy. Biology and Philosophy 17:567–590.

Associate Editor: Bernard J. Crespi