BEHAVIORAL AND BRAIN SCIENCES (2012) **35**, 1–59
doi:10.1017/S0140525X11000069

Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate

Francesco Guala
Department of Economics, University of Milan, 20122 Milan, Italy
francesco.guala@unimi.it
http://users.unimi.it/guala/index.htm

Abstract: Economists and biologists have proposed a distinction between two mechanisms – "strong" and "weak" reciprocity – that may explain the evolution of human sociality. Weak reciprocity theorists emphasize the benefits of long-term cooperation and the use of low-cost strategies to deter free-riders. Strong reciprocity theorists, in contrast, claim that cooperation in social dilemma games can be sustained by costly punishment mechanisms, even in one-shot and finitely repeated games. To support this claim, they have generated a large body of evidence concerning the willingness of experimental subjects to punish uncooperative free-riders at a cost to themselves. In this article, I distinguish between a "narrow" and a "wide" reading of the experimental evidence. Under the narrow reading, punishment experiments are just useful devices to measure psychological propensities in controlled laboratory conditions. Under the wide reading, they replicate a mechanism that supports cooperation also in "real-world" situations outside the laboratory. I argue that the wide interpretation must be tested using a combination of laboratory data and evidence about cooperation "in the wild." In spite of some often-repeated claims, there is no evidence that cooperation in the small egalitarian societies studied by anthropologists is enforced by means of costly punishment. Moreover, studies by economic and social historians show that social dilemmas in the wild are typically solved by institutions that coordinate punishment, reduce its cost, and extend the horizon of cooperation. The lack of field evidence for costly punishment suggests important constraints about what forms of cooperation can or cannot be sustained by means of decentralised policing.

Keywords: Cooperation; evolution; experiments; punishment; reciprocity

1. Introduction

Over the last two decades, research on human cooperation has made considerable progress on both the theoretical and the empirical front. Economists and biologists have proposed a distinction between two kinds of mechanism – "strong" and "weak" reciprocity – that may explain the evolution of human sociality. Reciprocity is, broadly speaking, a tendency to respond "nice" to nice actions and "nasty" to nasty actions when interacting with other players. Models of *weak* reciprocity require that reciprocal strategies be profitable for the agents who play them. *Strong* reciprocity models, in contrast, allow players to choose suboptimal strategies, and thus diverge substantially from the models of self-interested behaviour that are typically used by evolutionary biologists and rational choice theorists.[1]

The behaviour of strong reciprocators can be less than optimal in roughly two ways: On the one hand, strong reciprocators play cooperatively with cooperators, even though it would be more advantageous to exploit them (let us call it *positive* strong reciprocity). On the other, strong reciprocators are willing to punish defectors at a cost to themselves, even though it would be advantageous to simply ignore them (*negative* strong reciprocity). These two types of action constitute the "bright" and the "dark" side of reciprocity, so to speak.

Both sides of reciprocity may be necessary to sustain human cooperation. In a heterogeneous population, a small fraction of free-riders can drive positive reciprocators towards low levels of cooperation. Costly punishment in such circumstances may provide enough policing to preserve an environment where cooperation can thrive. To support this claim, strong reciprocity theorists have generated a large body of evidence concerning the willingness of experimental subjects to punish uncooperative free-riders at a cost for themselves. This evidence and its theoretical implications constitute the main topic of this article. Although positive reciprocity is at least as important for the mechanics of cooperation, it deserves a separate analysis and will not be discussed except briefly at the end.

FRANCESCO GUALA is Associate Professor in the Department of Economics at the University of Milan (Italy). He works primarily on the philosophical foundations of social science, using experimental and theoretical methods. He is the author of *The Methodology of Experimental Economics* (Cambridge University Press, 2005) and co-editor of *The Philosophy of Social Science Reader* (Routledge, 2011). In 2002 he was the recipient of both the International Network of Economic Method Prize and the History of Economic Analysis Award. In 2009 he has been awarded a special "anti-brain-drain" scholarship by the Italian Ministry of Higher Education.

 0140-525X/12 $40.00

I argue that the message of punishment experiments is far from clear. To dispel some confusion, I introduce a few preliminary distinctions between concepts (such as absolute and relative costs, symbolic and material, and coordinated and uncoordinated punishment) that are often conflated in the writings of reciprocity theorists. It turns out that experimental results can be interpreted in different ways, and that while some interpretations are empirically warranted, others are just unproven conjectures at this stage. The first purpose of this article is to clarify the methods used by economists and biologists and help the resolution of open issues in reciprocity theory.

I distinguish between a "narrow" and a "wide" reading of the experimental evidence. Under the narrow reading, punishment experiments are just useful devices to measure robust psychological propensities ("social preferences") in controlled laboratory conditions. Under the wide reading, they replicate a mechanism that supports cooperation also in "real-world" situations outside the laboratory. These two interpretations must be kept separate because cooperation outside the laboratory may be sustained by mechanisms that have little to do with those studied by experimental economists.

I shall argue that the wide interpretation can only be tested using a combination of laboratory data and evidence about cooperation "in the wild." Field evidence, however, brings bad news for strong reciprocity theorists. I will focus on two points in particular: First, in spite of some often-repeated claims, there is no evidence that cooperation in the small egalitarian societies studied by anthropologists is enforced by means of costly punishment. Second, studies by economic and social historians show that social dilemmas in the wild are typically solved by institutions that reduce the costs of decentralized punishment and facilitate the functioning of weak reciprocity mechanisms. The second goal of this article, then, is to survey relevant evidence from history and anthropology that economists are usually unfamiliar with, and which is sometimes misrepresented by reciprocity theorists.

The conclusions to be drawn from this exercise, however, are not entirely negative for strong reciprocity theory. I shall argue that costly punishment experiments may still be useful as measurement devices, to observe motives that would otherwise be difficult to detect outside the laboratory. Negative and positive reciprocity, moreover, may be governed by different mechanisms, and failure on one front does not imply failure on the other. Still, the lack of field evidence for costly punishment suggests important constraints about which forms of cooperation can or cannot be sustained by means of decentralised monitoring and policing.

2. Reciprocity and social cooperation

The problem of cooperation is one of the classic puzzles of social science and political philosophy. Following a tradition that goes back to Hobbes, social theorists have used the Prisoner's Dilemma to represent the problem of cooperation in a situation where each individual has an incentive to defect from the social contract and free-ride on the fruits of others' labour (Fig. 1). This is the "State of Nature" of classic political philosophy, where no player can trust the others to behave pro-socially.

	C	D
C	2, 2	0, 3
D	3, 0	1, 1

Figure 1. A Prisoner's Dilemma game. The usual conventions apply: The strategies of Player 1 are represented as rows, and those of Player 2 as columns. The first number in each cell is the payoff of Player 1; the second one of Player 2.

In the Prisoner's Dilemma game two players must choose simultaneously one of two strategies, Cooperate (C) or Defect (D). It is immediately obvious that mutual cooperation (CC) is more efficient than mutual defection (DD). The payoffs of the game, however, are designed in such a way that each player has an incentive to defect, regardless of what the other player does. If the other player cooperates, defection delivers three units of payoff instead of two; if the other defects, it guarantees one unit instead of nothing. But this reasoning should lead both players to defect: In game-theoretic jargon, mutual defection (DD) is the only Nash equilibrium in the one-shot Prisoner's Dilemma.

A Nash equilibrium is a set of strategies (one for each player in a game) such that no one can do better by changing her strategy unilaterally. Nash equilibria are self-sustaining, or self-policing, in the sense that they are robust to individual attempts to gain by deviating from the current strategies (because, quite simply, no such gains are possible). It seems highly desirable that social institutions should be Nash equilibria, for they would be robust to exploitation and the constant threat posed by individual greed. "Cooperate" in the Prisoner's Dilemma is a prototypical rule that would enhance social welfare if generally endorsed by the members of the group. It is not, however, a stable institution, for it is not a Nash equilibrium of this simple game. Although mutual cooperation (CC) is more efficient than mutual defection, it is strictly dominated and will not be played by rational selfish individuals. If the social contract game were a one-shot Prisoner's Dilemma, then a population of rational players would never be able to pull themselves out of the war of all against all.

For many social scientists, the puzzle of cooperation is just an artefact of the peculiar behavioural assumptions of standard economic theory: Surely only selfish economic agents defect in dilemma games, while the rest of us – "the folk" – can do much better than that. But this view is simplistic. Far from being an arbitrary assumption, the self-interest principle is well-rooted in evolutionary theorizing. Indeed, cooperation is in many ways more puzzling from a *biological*, than from an economic point of view.

"Biological altruism" denotes any behaviour that increases the chance of survival and reproduction of another (genetically unrelated) organism, at the expense of the altruist's direct fitness. Biologists have known for decades that the problem of biological altruism is structurally similar to a social dilemma game in economists' sense (Axelrod & Hamilton 1981; Dawkins 1976; Trivers 1971). An organism that does not help but receives help from others will produce on average more offspring, spreading its "selfish" genes more efficiently than its altruistic

fellows. Altruists (i.e., organisms playing C-strategies) should be washed out by the forces of natural selection, leaving only self-interested players behind.

But Homo sapiens' spectacular success, in fitness terms, surely has something to do with social cooperation. So the puzzle remains. According to a prominent tradition in economics and biology, the solution lies in the concept of *reciprocity*. Reciprocity is a human propensity to respond with kindness to kind actions, and with hostility to nasty actions. Its logic is encapsulated in different cultures by Golden-Rule principles such as "Do to others what you would like to be done to you" and "Hurt no one so that no one may hurt you."

Reciprocity theory bloomed in the 1970s when game theorists and theoretical biologists almost simultaneously began to study the properties of conditional strategies in repeated games.[2] Robert Axelrod's (1984) tournaments are perhaps the best-known setting of this kind. Axelrod experimented with artificial players competing in a series of repeated dilemma games. Famously, a strategy called "Tit-for-tat" emerged as the winner in these tournaments. Tit-for-tat is a rudimentary rule of reciprocity, offering cooperation at the outset and then copying whatever move one's partner has made in the previous round. In spite of several limitations (Bendor & Swistak 1995; 1997; Binmore 1998, Ch. 3), Axelrod's simulations convinced many scholars that reciprocity can sustain cooperation in the long run, and that pairs of reciprocators are more efficient producers of resources than selfish free-riders.

This insight had a precursor in the biological concept of "reciprocal altruism" (Trivers 1971), the idea that what seems altruistic in the short run might actually be self-serving in the long term. Organisms that help others may be indirectly maximizing their own fitness, if their help is going to be reciprocated in the future. To capture the self-serving aspect of cooperation, I classify these approaches under the umbrella of "weak reciprocity" theory, and distinguish them from alternative ("strong") models that – paraphrasing Trivers – are not designed to "take the altruism out of altruism" (see sect. 3).

Axelrod's (1984) and Trivers' (1971) findings are consistent with a general game-theoretic result known as the *folk theorem* (Fudenberg & Maskin 1986; Fudenberg et al. 1994). Informally, the folk theorem says that any strategy guaranteeing at least as much as the worst payoff that can be inflicted by the other player is a Nash equilibrium of an indefinitely repeated game. In the repeated Prisoner's Dilemma, a partner who does not reciprocate can be punished by withdrawing cooperation, a mechanism known as "trigger strategy" in game theory. Suppose that in the first round, I play cooperate and you play defect. From the second round, I can "punish" you by defecting, ensuring that your future stream of payoffs is not greater than one unit per period. Because mutual cooperation would guarantee an expected average payoff of two units, you are better off cooperating right from the start. The threat of defection makes mutual cooperation attractive, if the shadow of the future is long enough to make it worthwhile.

The folk theorem carries good and bad news for evolutionary social theory. The good news is that in an indefinitely repeated game, cooperation is sustainable using trigger strategies that punish deviation from cooperative behaviour and cancel the advantages of defection. The bad news is that *infinitely* many strategies are Nash equilibria of this sort. Tit-for-tat is only one among many equilibria in the infinitely repeated Prisoner's Dilemma. Consider a strategy profile such as "I cooperate on Monday, Wednesday, and Friday, and you cooperate on Tuesday, Thursday, Saturday, and Sunday." Using the matrix of Figure 1, such a profile delivers an average payoff of 1.71 to me, and 1.28 to you. Because it is better than the worst penalty I can inflict (by withdrawing cooperation) if you do not follow it, it is a Nash equilibrium of the indefinitely repeated game. But like many other strategy profiles, it is not equitable (in many ways, in fact, it is intuitively unfair).

How can we identify, among all the possible equilibria, the ones that will be actually played? Communication can certainly improve coordination among organisms – such as humans – who have the capacity to exchange signals. Moreover, it is possible that selection drives out inefficient signals and their respective equilibria in the long run. The idea is that richer, more productive societies may outperform less efficient ones and replace them by absorption, extinction, or a combination of both. This is known in theoretical biology as the process of *group selection*. Although it came in disrepute during the 1970s, the idea that selection can operate at group level has been rehabilitated and is now widely used to explain processes of social evolution (Bergstrom 2002; Boyd & Richerson 1990; Wilson & Sober 1994). If homogeneous groups of conditional cooperators are more efficient, in the long run they should be able to outperform homogeneous groups of free-riders trapped in suboptimal equilibria. This result, again, holds only under certain restrictive conditions (for group selection to operate, for example, groups must be relatively stable and impermeable to immigrants carrying different traits), but it gives us the beginning of an explanation of the evolution of social cooperation.

3. Strong reciprocity

In weak reciprocity theory withdrawing cooperation is a strategy of self-defence that damages the free-rider but benefits the reciprocator. Weak reciprocity mechanisms therefore appeal to individuals' self-interest (as well as foresight). The folk theorem, for example, does not require that we relax the standard assumptions of self-interest and rationality of neoclassical economic theory. Similarly, Trivers' (1971) "reciprocal altruism" is not a disinterested form of altruism: A missed opportunity to exploit others' cooperation *now*, to be sustainable, must be fully repaid by mutual cooperation in the future.

Explaining cooperation by individual self-interest, however, comes at a price. Three conditions limit the application of weak reciprocity mechanisms to a rather narrow set of circumstances: First, the shadow of the future must be long enough, in an objective *and* subjective sense: the players must not discount future payoffs too heavily, otherwise the temptation to defect will be strong regardless of the future stream of gains from cooperation. Second, the number of players must be small, so that monitoring cooperation is relatively easy, and the withdrawal of cooperation does not damage too many cooperators.

Third, information in the group must circulate freely and without error, for otherwise the threat of punishment will be ineffective. When some of these conditions do not apply, the folk theorem holds only for unrealistically high values of the other parameters (Fudenberg et al. 1994).

These limitations, according to some critics, make the folk theorem a poor tool for the analysis of social cooperation (Gintis 2006; 2009). Discounting future gains is a well-established fact of human psychology; in large modern societies, moreover, one-shot encounters with unrelated strangers are ubiquitous, and information is rarely transparent. So, the critics argue, we need a kind of reciprocity that is able to sustain cooperation where weak reciprocity cannot reach and folk-theorem mechanisms fail.

Strong reciprocity theory is the result of collaboration between experimental economists, game theorists, anthropologists, and theoretical biologists interested in the evolution of human cooperation. Samuel Bowles, Herbert Gintis, Ernst Fehr, Robert Boyd, and Peter Richerson are its best-known advocates, but many other social scientists and biologists have contributed to its success (Gintis et al. 2005; Henrich et al. 2004). The theory departs from the classic approach by modelling "strong reciprocators," who, unlike weak ones, are not solely concerned about future gains. Strong reciprocators cooperate because they feel it is the right thing to do, and they are ready to punish defectors at a cost. Punishment is not merely withdrawal from cooperation, but involves the subtraction of resources from free-riders. Since taking resources away requires an active effort or risk, punishers pay a fee that is subtracted from their earnings. The act of punishment results in an immediate reduction of welfare both for the punisher and for the punished individual.

Strong reciprocity nevertheless has some important advantages compared to its weak cousin. Costly punishment ensures that defectors do not enjoy the fruits of free-riding. Free-riders, moreover, are punished by strong reciprocators even in one-shot games and when the future is heavily discounted. Strong reciprocity thus can potentially support cooperation even in large groups, where repeated encounters are rare or unlikely, and interactions with strangers are common. Costly punishment changes radically the incentives of free-riders, without affecting the other cooperators in the group.

The logic of punishment, however, takes the form of a "second-order" social dilemma: Sanctions are a public good that benefit all cooperators in the group, but are paid by the punisher only. In principle, everybody would like free-riders to be punished, but would prefer that somebody else do it. A plausible hypothesis is that the second-order dilemma is solved by automatic mechanisms – such as emotions, internalized norms, or social preferences – that bypass strategic considerations and trigger actions that would be avoided by a rational selfish calculator (Frank 1988; Hirshleifer 1987). But could these mechanisms have survived Darwinian selection?

Simulations suggest that cooperative strategies can evolve in favourable conditions (Bowles & Gintis 2004; Boyd & Richerson 1992; Boyd et al. 2003; Gintis 2000; Henrich & Boyd 2001). These conditions include a certain degree of behavioural homogeneity within groups, trait diversity across groups, and selection mechanisms that grant a higher survival rate to members of more cooperative groups. Notice that the problem of multiple equilibria is more severe for strong reciprocators because costly punishment can support an even wider range of equilibria, including equilibria that are not welfare- or fitness-enhancing for the group (Boyd & Richerson 1992). A community, for example, may be prevented from adopting a set of beneficial strategies, simply because they depart from what is considered "correct" behaviour by an aggressive gang of moralistic punishers. In such conditions, evolution arguably must play an even more important role in the process of equilibrium selection, than in classic weak reciprocity models.

4. Costly punishment in the laboratory

The picture of human motives painted by strong reciprocity theory is intuitively appealing – but is it empirically accurate? Since the 1980s the strongest evidence in its favour has come from laboratory experiments, and therefore we will have to examine data and experimental designs in some detail. Although the experimental literature is already large – and constantly growing – it is driven by a set of core results and robust patterns, which are the main focus of this article.

Experimental economists' interest in costly punishment derives from the analysis of a simple bargaining setting known as the Ultimatum Game (Güth et al. 1982). The Ultimatum Game is the simplest sequential bargaining situation that one can think of: Two players have the opportunity to share a sum of money (say, 10 dollars). Player 1 has the advantage of making the first offer. This introduces an important power asymmetry: Player 2 at this point can only accept or reject. If Player 2 accepts, the two players earn the proposed amount; if Player 2 rejects, they walk out with nothing. The unique subgame perfect equilibrium of the Ultimatum game is for Player 1 to offer as little as possible (one dollar, for example), and for Player 2 to accept because a dollar is better than nothing. The Ultimatum Game in theory should give rise to very inequitable distributions of resources.

When the Ultimatum Game is played for real, however, fair allocations figure prominently. Experiments in North America and West Europe result in average offers between 30% and 40% of the endowment, and a mode at the 50–50 split. Unfair offers (of 30% or less) are rejected about half of the time (Camerer 2003). A common interpretation of this behaviour is in terms of strong negative reciprocity: People are willing to pay a cost to punish offers that they perceive as unfair, even though they are not going to meet the offender ever again. By so doing, they fulfill a useful social function, for unfair players learn what is expected of them, and conform to the prevailing norm in future encounters.[3]

The insight of the Ultimatum Game can be extended to other game-theoretic settings. In a widely cited series of experiments, a group of economists led by Ernst Fehr have studied the effect of punishment on cooperation in public goods and other social dilemma games (e.g., Falk et al. 2005; Fehr & Fischbacher 2004; 2005; Fehr & Gächter 2000a; 2002; de Quervain et al. 2004); this approach was pioneered by Yamagishi (1986) and

Ostrom et al. (1992). The classic dilemma situation is modified adding a second stage, in which cooperators can punish free-riders and destroy what they have illicitly gained. Punishment comes at a cost, however, in the form of a fee paid by the subjects who voluntarily engage in this sort of policing.

Adding the punishment phase radically changes the game. Take the simplest case of a two-player Prisoner's Dilemma, as shown in Figure 2: The matrix in Figure 2a is turned into a more complex game by the addition of an extra strategy P in the second stage for the cheated player, as in Figure 2b.[4] Suppose, for example, that Row defects while Column cooperates. The outcome of the first stage of the game is (3, 0), but in the second stage Column is given the opportunity to move unilaterally from (3, 0) to (0, −1). This option is strictly dominated, and should not be chosen by a rational self-interested player. Yet, if Player 2 manages to convince Player 1 that she will play P, she will effectively transform the Prisoner's Dilemma into a coordination problem such as that of Figure 2c, where mutual cooperation (CC) is a Nash equilibrium of the game, and a Pareto-efficient one as well.

This is apparently what happens in standard punishment experiments with public goods games: in spite of the fee, many people are willing to sanction, and their threat is credible enough to raise cooperation to high levels (Fehr & Gächter 2000a; 2002). This result holds both when the game is played repeatedly by the same players (for a finite number of rounds), and when the membership of the group changes at every round. Costly punishment is administered even in one-shot games (Gächter & Herrmann 2009; Walker & Halloran 2004) and by "bystanders" or "third parties" – that is, when the potential punishers are not themselves the victims but have merely witnessed exploitative behaviour (Fehr & Fischbacher 2004) – although in such cases it does not always raise the average level of cooperation.

Recent studies with brain imaging have provided further insights about the psychological and neural mechanisms implicated in such behaviour. Costly punishment seems to be partly triggered by an impulsive negative reaction against injustice (Sanfey et al. 2003) and partly motivated by the sheer pleasure of punishing social deviants (de Quervain et al. 2004). Building on this evidence, strong reciprocity theorists have argued that reciprocal motives are robust enough to be represented as "social preferences" governing individual behaviour across a variety of decision tasks. Although the formal representation of reciprocity raises a number of difficult technical issues, various models have been proposed in the game theory literature, and probably even more will appear in the future (see Falk & Fischbacher 2005 for a survey).

(a)

	C	D
C	2, 2	0, 3
D	3, 0	1, 1

(b)

	C	D	P
C	2, 2	0, 3	
D	3, 0	1, 1	0, -1
P		-1, 0	

(c)

	C	D
C	2, 2	-1, 0
D	0, -1	1, 1

Figure 2. Transformation of a Prisoner's Dilemma (a) into a coordination game (c), by the addition of a punishment option (b).

5. Two interpretations of punishment experiments

Costly punishment is robust to replication, a real experimental phenomenon that can teach us something about the mechanics of cooperation. And yet, it is not clear *what* it does teach, exactly. In this section I will argue that the success of strong reciprocity theory derives in part from equivocating two possible readings of punishment experiments – "narrow" and "wide" – which have different epistemic statuses and implications. While the narrow reading is unobjectionable, it will turn out that the wide one is currently little more than a conjecture. Since its popularity is partly due to its conflation with the narrow (and empirically warranted) interpretation, it is important to distinguish them clearly before we proceed.

According to the *narrow* interpretation, punishment experiments open an interesting window on psychological motives and reactions to violations of social norms. In a review aimed at advertising punishment experiments among non-economists, for example, Colin Camerer and Ernst Fehr write that "the purpose of this chapter is to describe a menu of experimental games that are useful for measuring aspects of social norms and social preferences" (Camerer & Fehr 2004, p. 55). The punishment design seems to be motivated primarily by methodological concerns, rather than by realism. Similarly, according to Fehr and Schmidt,

> All these games share the feature of simplicity. Because they are so simple, they are easy to understand for the experimental subjects and this makes inferences about subjects' motives more convincing. (Fehr & Schmidt 2006, p. 621)

Under this interpretation, punishment mechanisms are useful *methodological devices to observe social preferences.* (I use the term "preference" broadly, to cover all sorts of dispositions including desires, emotions, and feelings; on the use of experiments as measurement devices, see Guala 2008). This narrow reading is uncontroversial: As far as I am aware, nobody denies that punishment experiments can be used to learn about human attitudes towards cooperation in the lab. But the narrow interpretation does *not* imply that costly punishment sustains social cooperation in the real world. Costly punishment is just the experimenter's way of turning unobservable attitudes and dispositions ("preferences") into observable and quantifiable experimental variables.

The *wide* interpretation of punishment experiments is bolder: Punishment mechanisms are not just measurement devices, but replicate in the laboratory the same processes that support cooperation in the real world. There is no doubt that strong reciprocity theorists interpret their experiments in this wide sense, to support a general account of cooperation based on costly punishment mechanisms. In one of the seminal papers in this literature, for example, Fehr and Gächter claim that "in our view punishment of free-riding also plays an important role in real life" (2000, p. 993). Influential anthropologists Boyd and Richerson add that:

> Fehr's experiment suggests that some of the neighbors watching us take sadistic pleasure in punishing our transgressions, or

> at least feel obligated to exert considerable effort to punish. Worrying about what unselfishly moralistic neighbors will do is an entirely reasonable precaution for humans. (Richerson & Boyd 2005, p. 220)

Following the anthropologists' lead, Camerer and Fehr suggest that costly punishment sustains cooperative practices such as food sharing in small groups of hunter-gatherers:

> Reciprocity, inequality aversion, and altruism can have large effects on the regularities of social life and, in particular, on the enforcement of social norms. … For example, if many people in a society exhibit inequality aversion or reciprocity, they will be willing to punish those who do not share food, so no formal mechanism is needed to govern food sharing. Without such preferences, formal mechanisms are needed to sustain food sharing (or sharing does not occur at all). (Camerer & Fehr 2004, p. 56)

According to Fehr and Fischbacher:

> This kind of punishment [observed in the laboratory] mimics an angry group member scolding a free-rider or spreading the word so that the free-rider is ostracized – there is some cost to the punisher, but a large cost to the free-rider. (Fehr & Fischbacher 2005, p. 169; see also Camerer & Fehr 2004, p. 68, for an almost verbatim repetition of this statement)

The narrow and wide interpretations of punishment experiments correspond roughly to two levels of *validity* of experimental results that are sometimes distinguished in the methodological literature in psychology, economics, and biology (Bardsley et al. 2009; Guala 2005; Steel 2007). According to this distinction, an experimental result is *internally valid* when the experimenters have correctly inferred the causal factors or mechanisms that generate data in a particular laboratory setting. Identifying data-generating processes in the lab, however, is rarely the ultimate goal of experimenters in the social sciences. Researchers typically want to find out about variables and processes that play an important role in a class of *non*-laboratory phenomena of interest (phenomena "in the real world," as they sometimes put it). The wide interpretation makes the additional claim that experimental results can be extrapolated to explain cooperation in some class of non-laboratory conditions, and so it amounts to an *external validity* inference.

Notice that the result of every well-designed experiment is valid, trivially, in all non-laboratory circumstances that replicate exactly the experimental conditions. But since the point of running controlled experiments is to create conditions that cannot easily be found in nature, where specific theories can be rigorously tested and new hypotheses investigated, the application of experimental results typically involves an external validity inference or generalization. This generalization requires extra evidence to be sustained, and the quality of this evidence, as we shall see, is very questionable in the case of costly punishment.

6. Experiments in the field

Costly punishment is used explicitly to explain cooperation in large societies, where one-shot encounters are common and information is poor. This may suggest that the punishment story accounts for a real-world phenomenon and is not just the artefact of a peculiar experimental setting. But this conclusion would be too hasty, for disagreement between the weak and strong reciprocity camps begins at the level of the phenomenon to be explained. Critics of the costly punishment story usually hold that one-shot cooperation among strangers in large-scale societies does not take place (except sporadically and unsystematically): The limits of the folk theorem are the limits of spontaneous cooperation. Outside the boundaries of the family, the small circle of a local community, or the long-term relationships we cultivate with business partners, we need *other* incentives (such as those provided by centralized policing) to prevent exploitation, free-riding, and abuse of power (e.g., Binmore 2005, p. 82). Weak reciprocity theory, in other words, draws different boundaries for spontaneous cooperation, and cannot be blamed (without begging the question) for its presumed "failure" to explain a phenomenon that by its own light may well not exist.

The key source of disagreement, then, is spontaneous cooperation *outside* the lab. Supporters of strong reciprocity sometimes seem to claim that costly punishment has been observed in the field, which obviously would resolve the issue of validity at once. But such a claim, again, trades on ambiguity. Costly punishment has indeed been observed across subject pools in several developed countries, as well as in Ultimatum and Public Goods experiments run in small-scale societies (Henrich et al. 2004; Herrmann et al. 2008; Marlowe et al. 2008). But none of these studies investigates behaviour in a natural setting or amounts to a *natural field experiment* as the term is used in economics.

Harrison and List (2004) distinguish between *artefactual* and *natural* field experiments. A natural field experiment successfully manipulates one variable of interest in an environment that is otherwise left as much as possible unaffected by the experimenter. Ideally, the subjects should be unaware that they are participating in an experiment, and select their responses from a menu of strategies that they normally use in their everyday lives. Artefactual experiments, in contrast, differ from conventional laboratory studies only with respect to the sample of subjects, which is drawn from the target population instead of some more convenient pool (e.g., a population of African bushmen, as opposed to university undergraduates, if we are studying cooperation in small-scale societies). The strategic setting and the framing, however, are imposed by design instead of mirroring a realistic decision-making environment (whether the experiments are performed in a university lab or in a hut in the African forest is irrelevant). So-called field experiments with punishment are artefactual in this sense, for they involve situations that are probably quite unfamiliar to the decision makers, and as we shall see, they do not reproduce the full menu of strategies that are available in the dilemmas of cooperation that people face in everyday life. In fact, it would be more appropriate to speak of "experiments in the field" in this case, rather than "field experiments."

This is not merely a terminological quibble. Artefactual designs raise serious issues for the wide interpretation of punishment experiments. As the terminology suggests, these experiments are more likely to generate experimental artefacts than natural ones. This does not mean, of course, that they are useless. On the contrary, they are extremely helpful because they guarantee a higher degree of control on the environment and allow the elimination of potentially confounding variables that may elude

control in the field. It does not mean that experimental phenomena such as costly punishment are somewhat *unreal* either. A phenomenon may be real *and* artefactual – a real experimental effect generated in circumstances that do not mirror those naturally found in the natural or social world (Hacking 1988). As we shall see, there are good reasons to believe that costly punishment is a "real artefact" in this sense of the term: Artefactual insofar as it is produced by the specific experimental procedures, but nevertheless real because it does take place in a limited range of (laboratory-like) conditions.

7. Repetition and evolutionary scale

External validity objections can hinder scientific progress when they are meant to raise sceptical doubts about the use of experiments generally. But external validity worries are inescapable and indeed useful when addressed to the specific details of an experimental design, for they help establish the reliability of specific inferences from the laboratory to field settings (cf. Bardsley et al. 2009; Guala 2005; Starmer 1999). It is in the latter spirit that one must ask whether costly punishment is an artefact of the experimental setting that economists implement in their laboratories.

One major external validity problem has to do with *scale:* Both strong and weak reciprocity models describe behaviour on an evolutionary time-scale and are not primarily intended to capture choices in experimental games that last only for a short time (Binmore 1998; 2005; Ross 2006). Of course, there is no reason to expect that what evolves in the long run is similar to what we observe in the short run of experimental games. When people play Ultimatum Games in the laboratory, for example, they may bring with them norms and heuristics that help coordination in their everyday dealings. Such dealings are often in the form of indefinitely repeated games, where egalitarian splits can be sustained by weak reciprocity mechanisms. The behaviour observed in the laboratory thus may be a misapplication, in an unfamiliar setting, of a heuristic rule that works well (and was selected for) in the larger but more familiar games that we play in real life. If the experimental games were repeated long enough, however, out-of-equilibrium strategies would be eliminated by evolutionary forces and learning, until behaviour approaches a rational equilibrium (cf. Binmore 1998; 1999; 2006; Burnham & Johnson 2005; Hagen & Hammerstein 2006; Trivers 2004).

This argument sounds plausible, but unfortunately it is inconclusive. To begin with, it is easy to retort that prosocial behaviour in settings such as the Ultimatum Game is remarkably robust even when the games are repeated for several rounds (e.g., Cooper & Dutcher 2009; Roth et al. 1991). The rate of costly punishment has been observed to increase, rather than decrease, after as many as 50 rounds of play (Gächter et al. 2008). If strong reciprocity "misfires" in finitely repeated games, it does so systematically enough to be of theoretical interest for social scientists (Richerson & Boyd & 2005, p. 220), since what happens in the very long run is irrelevant for the many short-run games that we play in the lab and in real life. Next, there is evidence that experimental subjects distinguish between one-shot and finitely repeated games, and modulate their strategies accordingly (Fehr & Fischbacher 2002; 2005; Gintis et al. 2003). To insist that they do not understand the difference between finitely and indefinitely repeated games, as some critics do (e.g., Binmore 1999; 2006), therefore seems arbitrary and unjustified.

These replies are powerful, and the critics of strong reciprocity theory are wrong to insist on this line of argument. From a logical point of view, one can keep asking whether costly punishment would survive hundreds or thousands of repetitions. (How many times can you get angry in an indefinitely repeated Ultimatum Game?) And yet, this challenge in itself does not lead to any new testable proposition: It belongs to the class of sceptical challenges to experimentation that bring the discussion to a halt, unless new evidence is offered in support.

Complemented with new data, in contrast, external validity worries can become a powerful engine for scientific progress – they can be used to make interesting predictions that are tested empirically. It is in this constructive spirit that we must look for field data concerning costly punishment. To assess the wide interpretation of punishment experiments, we must study "richer" situations, where decision makers can choose from the full range of strategies that are customarily available in everyday life. Natural field experiments are richer just in this sense. But since there are no natural field experiments on costly punishment, we ought to look for relevant data elsewhere. The next four sections (8 to 11) review non-experimental evidence that is seldom discussed by theorists on either side of the controversy, that is, ethnographic data from anthropology, a source that is often cited by reciprocity theorists but never analysed in much depth. I shall return briefly to laboratory data in section 12, while section 13 deals with historical evidence concerning common pool institutions.

8. Costly punishment in small societies

The Leviathan is a relatively recent invention. During most of their evolutionary history, *Homo sapiens* probably lived in small egalitarian bands without a centralised leadership. The head of each family enjoyed a high degree of autonomy in decision-making, and even the most authoritative men in the band could only persuade – never force – others to follow a certain course of action. In the words of Marshall Sahlins:

> The indicative condition of primitive society is the absence of a public and sovereign power: persons and (especially) groups confront each other not merely as distinct interests but with the possible inclination and certain right to physically prosecute these interests. Force is decentralized, legitimately held in severalty, the social compact has yet to be drawn, the state nonexistent. So peacemaking is not a sporadic intersocietal event, it is a continuous process going on within society itself. (Sahlins 1972/1974, pp. 186–87)

The small-scale societies of hunter-gatherers, horticulturalists, and nomadic pastoralists that have been studied extensively by anthropologists are probably the last remnants of these ancient acephalous social orders based on spontaneous cooperation. Although strong reciprocity theorists say that their models explain the emergence and

maintenance of cooperation in small egalitarian societies, they provide surprisingly thin evidence in support.

According to Bowles and Gintis (2002, p. 128), for example, "studies of contemporary hunter-gatherers and other evidence suggest that altruistic punishment may have been common in mobile foraging bands during the first 100,000 years or so of the existence of modern humans." In support of this claim, however, they cite a study (Boehm 1999) that does *not* endorse a costly punishment account of human sociality. Richerson and Boyd (2005, p. 219) write that "in small-scale societies, considerable ethnographic evidence suggests that moral norms are enforced by punishment." Among their references, however, one finds only two ethnographic surveys, a laboratory experiment, and a study of dominance that do *not* support the costly punishment story (cf. Richerson & Boyd 2005, p. 280, n. 60).

Most of Richerson and Boyd's (2005) case is, in fact, based on Fehr and Gächter's (2000a; 2002) experiments. Fehr and his colleagues state that "private sanctions have enforced social norms for millennia, long before legal enforcement institutions existed, and punishment by peers still represents a powerful norm enforcement device, even in contemporary Western societies" (Spitzer et al. 2007, p. 185). "The prominent role of such peer punishment" is reported as an established fact, even though their bibliography refers only to a laboratory experiment (Fehr & Gächter 2002), an evolutionary model (Boyd et al. 2003), and a survey of ethnographic evidence that – again – does not support a costly punishment account of the evolution of cooperation (cf. Sober & Wilson 1998, pp. 166–68).

The costly punishment account of cooperation in small societies, then, seems to lack a solid base of ethnographic evidence. This is not surprising, for as we shall see, the available data are scarce. Before we look at the data more carefully, however, it is worth asking what kind of evidence would support the strong reciprocity story. Notice that all the aforementioned quotes tend to conflate costly punishment with punishment in general. But while there is no doubt that sanctions are crucial for the maintenance of social order, it is by no means obvious that they are costly for those who administer them. This is an important point that is often overlooked, or perhaps willingly confused in the literature: The very definition of strong reciprocity calls for evidence of *material* and *costly* punishment behaviour in field settings:

> [Strong] Reciprocity means that people are willing to reward friendly actions and to punish hostile actions *although the reward or punishment causes a net reduction in the material payoff of those who reward or punish.* (Camerer & Fehr 2004, p. 56, emphasis in the original)

More precisely, there are two kinds of cost that are relevant for our purposes. One is the *absolute* cost, the fee paid by an individual (in material terms) to punish a free-rider. The other one is the *relative* cost of punishment, that is, the difference between the net benefit of the punisher and the benefit of the other group members who choose not to punish. Absolute and relative costs must be kept separate because they raise different problems for different theoretical perspectives. When sanctions are costly in absolute terms, punishment cannot be explained using models based on self-interested motivation. If the cost is compensated by a positive benefit (to the punisher), in contrast, punishment is consistent with self-interest but potentially problematic from an evolutionary point of view. There may still be a relative cost in fact, and individual selection may work against the punisher (non-punishers may be advantaged in fitness terms, in other words).[5] But it is also possible that the relative cost is nil, because the costs are spread in such a way that everybody carries an equal share of the overall burden. In such a case, punishment would not be selected against within the group.[6]

Keeping these concepts in mind, we ought to ask two questions: Does punishment in small-scale societies involve an absolute cost? If so, is the cost borne by a single individual, or is it distributed across group members in such a way as to minimize the relative cost? Answering is not easy, because most of the evidence of punishment in small societies is anecdotal, and quantitative data regarding the frequency, intensity, and effect of material punishment are scarce. Another related problem is that the benefits from punishing a free-rider are often delayed, and even when we observe an immediate cost, we can rarely rule out that it will not be recouped at a later time. The most cooperative and popular members of a group, for example, may have easier access to sexual mates, an incentive mechanism that only bears fruit in the medium-long term of a reproductive cycle (Hawkes 1993).

Notice that for this reason cooperation in small societies does not constitute a very good test-case for strong reciprocity theory. Most interactions between the members of small societies take the form of an indefinitely repeated game, with relatively high monitoring and circulation of information. These interactions, far from being anonymous as in most laboratory experiments, rely crucially on reputation and trust (Wiessner 2009). This does not mean that such cooperation should be interpreted by default in weak reciprocity terms, of course; but it does mean that a priori the costly punishment story does not enjoy any advantage over its rival. Because cooperation in small societies is not mysterious or impossible from a weak reciprocity perspective, we ought to know more about the mechanics of coercion as it is described in the ethnographic literature.

9. Sex and death

Christopher Boehm has systematically surveyed and classified the ethnography on punishment and norm-violation. Boehm's (1999) work is the main source of empirical evidence for Sober and Wilson (1998), who in turn are widely cited by strong reciprocity theorists in spite of the fact that they do not support a costly punishment account of cooperation. Along the chain of citations Boehm's core message seems to have been lost.

Sanctions are ordered by Boehm (1999) on a scale that goes from ridicule, gossip, and verbal reproach, up to social ostracism and eventually homicide. Homicide is obviously the harshest and, because of the risk of retaliation, potentially the most expensive form of punishment. In relative terms, however, it is not rare. The view of primitive peoples as largely pacific has been abandoned by anthropologists over the last half-century, as the accumulation of statistical data has revealed a level of

endemic violence that is much higher than in most large sedentary societies (e.g., Chagnon 1988; Knauft 1991). The majority of violent confrontations within the tribe nevertheless are caused by sexual conflict rather than violation of norms of economic cooperation (Knauft 1991), and the punishment of adulterers by jealous husbands accounts for a large share of murders (Chagnon 1968/1992, p. 187; Lee 1979, p. 377; Marlowe 2010, p. 192).

Following Trivers' (1972) theory of parental investment, adultery can be plausibly modelled as a Prisoner's Dilemma game with fitness payoffs, and jealousy as an adaptive solution. Jealousy is a strong emotion that triggers aggressive behaviour, bypassing complex calculations of cost and benefit that might otherwise deter from the punishment of philanderers. To establish that revenge is systematically costly, however, requires some tricky quantitative analysis. If the probability of getting killed or injured during a fight (i.e., of compromising one's fitness) is lower than the probability of deterring sexual free-riders from sleeping with one's partner in the future, then revenge triggered by jealousy may be advantageous from an evolutionary point of view. Punishment need not be expensive in the long run, for the punisher would recoup the costs – for example, by gaining a reputation of "fierceness" that could promote access to sexual mates in the future.

Unfortunately, the available evidence is mostly qualitative, and only suggestive. Cultures of fierceness seem more common among horticulturalists like the Yanomamö than among mobile hunter-gatherers who can resolve their conflicts by frequent splitting. This seems to point in the direction of weak reciprocity mechanisms that exploit the long horizon of cooperation. But lacking precise data, any reciprocity account of adultery cannot be more than a conjecture.

One point, nevertheless, emerges strongly from the ethnographic literature: The violence that stems from sexual competition, far from contributing to sociality, is actually a major threat to the survival of small societies. Chagnon (1968/1992, p. 188) notes, for example, that dyadic club fights among the Yanomamö have a tendency to quickly escalate, and unless the elders are able to control them, they usually result in group fission. There may be a direct causal link between the size of groups, the opportunity to engage in adultery, and the probability of fission, which acts as a powerful limit on social aggregation. Lee (1979, p. 397) similarly claims that "the fear of violence ... is a prominent feature of !Kung life," and the Kalahari bushmen have developed various means to keep violence under control. One of these is simply to live in small groups of tightly related kin.

Because violent punishment hinders, rather than promotes, sociality, several mechanisms are in place to moderate the effects of male aggressiveness. Sexual tensions are often displaced or unacknowledged, and to some extent adultery is simply tolerated. Interestingly for our purposes, punishment is even less common in the case of economic, rather than sexual, free-riding: In her study of "costly" punishment among the Ju/'hoansi, Wiessner (2005, p. 134) noticed that "none of the cases with negative outcomes [for the punisher] dealt with regulation of sharing or [economic] free-riding." Shirkers are for the most part just ignored, an attitude that does not seem to be in any way peculiar to the Ju/'hoansi (see, e.g., Marlowe [2010] on the Hadza).

It is also significant that violent revenge is rarely praised, as one would expect in a society that relies on costly punishment for its survival; on the contrary, the murderer is often considered "polluted" and in need of purification. Sometimes the murderer is ostracized (Mahdi 1986), and sometimes the killing of a murderer by the victims' relatives is tolerated (a practice that comes very close to an "execution," in a society without central authority – see Lee 1979, Ch. 13). This is very different from the picture painted by strong reciprocity theorists. Far from posing a second-order Prisoner's Dilemma problem, violent acts of revenge risk being far too common in small acephalous societies.

Punishment experiments thus give a misleading appearance of orderly justice to a process that, in most cases, would trigger feuds and eventually degenerate into anarchy and war. In the laboratory this eventuality is typically prevented by design, because in the majority of experiments free-riders cannot revenge the moralistic sanctions they have received. (Recall the empty cells in Figure 2b: in most punishment experiments it is not possible to respond P to P.) But in those few experiments where *counter-punishing* is allowed, approximately one quarter of the sanctions are revenged. Moreover, the positive effect of strong reciprocity vanishes, causing a reduction of cooperation similar to that observed in experiments without punishment. And on top of that, aggregate payoffs are among the lowest observed in experimental Public Goods games (Denant-Boemont et al. 2007; Nikiforakis 2008).

So there are probably good reasons why decentralised, spontaneous material punishment is so rare outside the laboratory. In modern states decentralised sanctioning is explicitly forbidden by law, and anti-social behaviour is curtailed in ways that minimize the risk of feuds. Retaliation is controlled by imposing a monopoly of state violence, and the cost of punishment is recouped by compensating "professional punishers" (e.g., policemen). In small societies, apart from cases of sexual conflict, homicide is used occasionally to resolve political issues, such as the rise of a bullying chief (Boehm 1999). However, it is typically administered by a *coalition* against an individual – that is, in a way that resembles the centralised punishment typical of large-scale modern societies. The formation of coalitions and coordinated punishment is an important mechanism that is beginning to attract the attention of reciprocity theorists, so I will come back to it later (in section 13). Before that, it will be instructive to explore other mechanisms that sustain cooperation in small societies where costly material punishment is rarely administered.

10. Gossip and symbolic sanctions

Homicide and overt physical aggression account for only a fraction of punishment episodes reported by ethnographers of small societies. When justice is not administered centrally, violations of norms are mostly dealt with by means of sanctions that affect the material welfare of the recipient only indirectly, and at the same time impose little or no costs on those who administer them. Some critics of strong reciprocity theory have rightly pointed out that the evolution of higher cognitive capacities in

humans has brought as a side-effect a dramatic reduction in the cost of anti-social sanctioning (Binmore 2005, pp. 82–84; Ross 2006, pp. 65–67). Going down Boehm's (1999) list, in fact, it is clear that most sanctions do not fit neatly the definition of costly punishment. Take verbal reproach and ridicule, for example. The process of symbolic punishment is quite different from that of material punishment: Whereas the former is non-invasive, the latter is not. While the latter encourages physical aggression, the former does it on a much smaller scale. And although inflicting material punishment is likely to infringe upon individual rights that regulate the life of a group (e.g., property rights), symbolic punishment does not.

In experiments, subjects are even willing to pay a fee to administer symbolic punishment (Carpenter et al. 2004). Although this confirms that they have a strong motivation to manifest disapproval of norm violations, it is not clear that any fee has to be paid in real life. The ethnography of norm regulation emphasizes that gossip and reproach are low-cost strategies. "Spreading the word" usually takes the form of spontaneous gossiping, the chit-chat that accompanies most activities of nomadic foragers (see, e.g., Marshall 1961; Dunbar 1996/1998). The primary function of this constant flow of information is the necessity to sustain trust and monitor others' activities, as in folk-theorem accounts of repeated cooperation. Even when it is used as a sanctioning device, however, gossip is a *collective* endeavour – an important point to which I will return later – and certainly nothing like an individualistic initiative that requires considerable investment of time or the subtraction of resources from other profitable activities. "Speaking up first" against a norm violator is often cited as a costly act in the strong reciprocity literature because of the risk of retaliation, but there are very cheap ways of circulating information and forming coalitions against individual group members.

In her in-depth study of the Chaldean community in modern Detroit, Natalie Henrich reports that direct reproach is used only to sanction relatively minor violations of social norms (such as garbage recycling), whereas serious issues are always dealt with by "behind-your-back" gossip (Henrich & Henrich 2007, pp. 147–50). Because of its potentially destructive effect on reputation, gossip is a very powerful enforcement mechanism and is particularly feared by Chaldeans, with the added advantage of protecting the punishers from the wrath of their target.

Polly Wiessner (2005) has made a systematic attempt to find evidence of costly punishment in the field, using ethnographic evidence collected among the bushmen of Botswana. Most of the punishment she reports is purely symbolic in character. Wiessner's conclusion is cautiously favourable to strong reciprocity theory, based on her estimate that 8% of observed punishment episodes had negative consequences for the punishers. Her definition of "negative consequence," however, is very broad, including cases like severed social relations and the loss of a group member through ostracism, which do not fit the proper definition of costly punishment. Wiessner does not distinguish between absolute and relative costs, but her discussion of the data suggests that both are very low in the case of economic dilemmas of cooperation. Even the risk of retaliation is extremely low: Physical confrontation, as a matter of fact, occurs in only 2% of the episodes recorded by Wiessner and never results in serious injuries (Wiessner 2005, p. 132). All in all, in a sample of 171 episodes, the statistical incidence of *material* cost for the punishers is close or equal to zero.

11. How pygmies punish free-riders

The next big step in the scale of sanctions reported by anthropologists is *ostracism.* Although descriptions of specific episodes are rare in the literature, ostracism figures prominently in Gurven's (2004) recent survey of the ethnographic record on food sharing, and experimental evidence confirms its efficacy in laboratory settings (Cinyabuguma et al. 2004; Page et al. 2005). Ostracism can be very damaging in material terms. Even though ostracized individuals or families usually join other groups, they lose ties with their kin and the protection that the latter provide. Among the Yanomamö studied by Chagnon (1968/1992), for example, leaving one's group entails leaving one's garden, and being dependent on the hosts for food for several months (the guests usually pay a "rent" in terms of women). Still, ostracism does not have to be costly: the exclusion of an individual or clan from the tribe usually takes place in such a way that no individual punisher has to bear the full "cost" of it. Ostracism can be so low-cost that it is often preferred to verbal reproach, especially in highly mobile societies: In such cases, it is not even necessary to expel the offender from the group – it is easier for the group to move elsewhere:

> When I ask the Hadza what they do if someone in a camp is being a slacker or being stingy, the most common answer is "we move away from them," rather than "we make them leave." They are averse to confrontations and solve most conflicts with others by moving. (Marlowe 2010, pp. 248–49)

To give an idea of how low-cost ostracism works, I will recount an episode reported by Colin Turnbull (1961) in his classic ethnography of the Mbuti pygmies in central Congo. Hunting is for the Mbuti a highly cooperative enterprise, involving all adult tribe members. Women work as beaters – they scare animals with screams and noises, pushing them towards an area of the forest that has been closed down using a line of nets. Once an animal is trapped in a net, it is speared by the nearest hunter who then "owns" the meat and is entitled to allocate it among the members of the hunting party, usually keeping the best parts for his own family. This technique requires the participation of several hunters, who must position themselves in an arc so as to close down a large area of the forest and act in concert to prevent the animals from escaping.

The band studied by Turnbull comprised several hunters, including a family headman named Cephu who was not well-liked and was already gossiped about in the group. Perhaps for this reason, Cephu occupied a peripheral location in the hunters' formation. This clearly put him at a disadvantage, since animals are more likely trapped in the middle sector of the line, and the hunters who occupy this sector end up with the largest share of the meat. On one particular occasion, the group had already killed a couple of preys when Cephu decided to abandon his position and, unseen, place his net in front of the other hunters. This is a typical free-riding strategy

in a social dilemma game: By changing location, Cephu increased the probability that the next animal caught in the trap would be speared by him, but at the same time he reduced the probability that an animal would be captured by the group at all.

On this particular occasion, Cephu's strategy was successful – he killed the first animal fleeing from the beaters – but did not go undetected. As Turnbull tells the story (1961, pp. 97–101), Cephu immediately became the victim of moralistic aggression by the whole group. While returning to the camp, several hunters began criticizing his conduct behind his back, with some of the youngsters ridiculing and insulting him amidst generalised laughter. This quickly escalated into a criticism of Cephu's anti-social behaviour in general, until an emergence meeting was called to resolve the matter once and forever. After a lame attempt to find an excuse, Cephu eventually tried to assert his right to occupy a better location in the line of nets, by virtue of his "chief" status. At this point, one of the other headmen simply and quietly invited him to leave the group, if he was too good and important to stay with the others on equal terms. This was sufficient to end the discussion. Here is Turnbull's description of subsequent events:

> Cephu knew he was defeated and humiliated. Alone, his band of three or four families was too small to make an efficient hunting unit. He apologized profusely, reiterating that he really did not know he had set up his nets in front of the others, and that in any case he would hand over all the meat. This settled the matter, and accompanied by most of the group he returned to his little camp and brusquely ordered his wife to hand over the spoils. She had little chance to refuse, as hands were already reaching into her basket and under the leaves of the roof of her hut where she had hidden her liver in anticipation of just such a contingency. Even her cooking pot was emptied. Then each of the other huts was searched and all the meat taken. Cephu's family protested loudly and everyone laughed at him. He clutched his stomach and said he would die; die because he was hungry and his brothers had taken away all his food; die because he was not respected. (Turnbull 1961, pp. 100–101)

Although this is clearly a case of material punishment, the punishment was certainly not very costly. First, the group made it clear that Cephu's conduct was considered unacceptable. The oral criticism was not just aimed at Cephu but was also for the benefit of the other members of the group, who were reassured about the balance of power. Then punishment was administered by a *coalition* against an individual (or a small clan) who would have no chance to counter-punish, and had no interest in escalating conflict.

Another interesting point is that the free-rider was punished by taking away his illicit gain. But, *pace* strong reciprocity theory, no wealth was destroyed, because the other families consumed what Cephu had caught. And even Cephu's punishment turned out to be not so harsh after all: Once peace had been restored, one member of the main group took some food to Cephu's hut to feed him and his family. At that point all animosity seemed to be gone, and Cephu participated in the feast with the rest of the group (Turnbull 1961, p. 101). (Cephu's clan, to be sure, abandoned the group later, to join another group of Mbuti hunters.)

Ostracism, as already mentioned, is described only rarely at this level of detail. Nevertheless, Cephu's story is representative of other episodes of moralistic aggression and ostracism reported in the anthropological literature (e.g., Briggs 1970; Boehm 1999, Ch. 3; for a survey, see Baumard 2010b). It shows that even cases that seem favourable (e.g., because they involve the subtraction of material resources) do not actually fit well with the explanatory framework of strong reciprocity theory. The expression "costly punishment" turns out to be a misnomer, because the punishment is inflicted in such a way as to keep both absolute and relative costs close to nil. Given the difficulty of obtaining a precise quantitative measurement, of course, one cannot rule out the costly punishment story with certainty. But it is fair to say that there is currently no evidence that cooperation is sustained by strong negative reciprocity in small societies. And whatever evidence there is, it rather points in the direction of cheap mechanisms like ostracism and coalitional punishment.

12. Cheap versus costly punishment in the lab

So why do people engage in costly punishment so enthusiastically in the laboratory? A plausible answer is that costly punishment is usually the only way for them to manifest their disappointment, and in any case punishers are protected by anonymity and by the rules of the experiment. But when they are given other options, subjects' behaviour changes: A handful of experiments have explored and compared the effects of different sanctioning techniques, ranging from purely symbolic (reproach) to purely material punishment. Evidence regarding the efficacy of symbolic sanctions is mixed, with some studies suggesting that reproaches backed by material punishment work best (cf. Janssen et al. 2010; Masclet et al. 2003; Noussair & Tucker 2005). If they are given the opportunity to choose, subjects prefer to support cooperation using a mix of symbolic communication, weak reciprocity, and the last-resort threat of material punishment (e.g., Ostrom et al. 1992; Rockenbach & Milinski 2006; Ule et al. 2009; Xiao & Houser 2005).

Most of these experiments, however, still ignore the problem of feuds and the anti-social effect of counter-punishment. There are to date only a couple of experimental studies that combine alternative ways of incentivising cooperation – including costly punishment – with the threat of counter-punishment. Nikiforakis and Engelmann (2010) find that strategies that could trigger lengthy feuds are avoided in the laboratory, and Dreber et al. (2008) show that in such circumstances people prefer to implement cheap strategies (i.e., withdraw cooperation) rather than costly punishment. This is sensible, because in the aggregate, feuds destroy more resources than they help create.

But even ignoring the problem of counter-punishment, "costly" punishment works only if it costs relatively little. Above a cost/impact ratio of 1:3, sanctions do not increase cooperation significantly (Egas & Riedl 2008; Nikiforakis & Normann 2008; Ohtsuki et al. 2009), and even low-cost punishment does not necessarily improve aggregate payoffs – in fact, it often reduces them. Although punishment pushes the rate of cooperation up, it also destroys resources. Janssen et al. (2010) report a strong positive effect on total revenue when communication is allowed,

and when it is matched with punishment, but not with punishment alone. Clearly this is deeply problematic, given the strong reciprocity theorists' emphasis on group selection.

An exception to this body of results is the discovery by Gächter et al. (2008) that costly sanctions can raise average earnings when the horizon of cooperation is very long (50 rounds). But notice that the game in this experiment involves repeated interactions with the same subjects, and efficiency increases because the long horizon makes the use of punishment almost unnecessary (there is more punishment in the final round than in the early part of the game, in fact). So, to sum up, costly punishment alone does not seem to be an efficient solution to social dilemmas in the laboratory, precisely in those conditions – such as one-shot interactions with strangers – where, according to strong reciprocity theorists, it would be most needed.

13. How common pool institutions sustain cooperation

I have discussed the ethnography of small societies in some detail because the behavioural scientists who are unfamiliar with this literature may be misled to believe that costly punishment is an established anthropological fact. But anthropology is not our only source of evidence concerning decentralised cooperation in the field, and small societies are neither the only nor the primary domain of application of strong reciprocity theory. Economic historians have studied extensively the spontaneous emergence of institutions for the management and preservation of public goods in complex societies. These studies emphasize that successful cooperative institutions solve social dilemma problems in ways that have little to do with costly punishment. Rather, they tackle the problem by removing the obstacles that prevent *non*-costly mechanisms from functioning.

We have a remarkable array of cases that can be brought to bear on this issue. I will briefly illustrate one example – the evolution of the *Carte di Regola* studied by Marco Casari (2007) in Northern Italy – that is representative of many similar institutions which have emerged spontaneously in different historical periods and in different parts of the world (see Ostrom 1990). All of them, as we shall see, have an important feature in common: They *artificially* create the conditions that, according to weak reciprocity accounts, make cooperation possible, but that for various reasons were *naturally* unavailable in the given circumstances.

The *Carte di Regola*, or "charters," are ancient written codes used by communities in the Trentino region in the northeast of Italy to regulate the exploitation of common pastures. The *Carte* were progressively introduced from 1200 until 1800, when they were eventually abolished by Napoleon. The charters were spontaneously adopted by single villages rather than imposed from above, and were aimed at preventing the over-exploitation of communal fields – a specific instance of Prisoner's Dilemma (or "common pool" problem) that has been studied in depth by historians (since McCloskey 1972). Using a database of more than two hundred villages, Casari (2007) has shown that the charters had a common structure and were aimed at removing precisely the obstacles that prevented weak reciprocity mechanisms from functioning well, even in isolated villages such as those in the Italian Alps. The *Carte*, to put it differently, made the application of (something like) the folk theorem possible.

A first set of charter rules enhanced the stability of local communities, by locking existing members in and preventing the entrance of opportunistic outsiders. This was done mainly by forbidding the sale of communal field rights (hence increasing the cost of leaving), and by requiring a supramajority consensus for the admission of new members. The only costless way of transmitting rights, then, was via inheritance through the head of the family, a mechanism that extended the horizon of cooperation across future generations and turned a finitely repeated game into an indefinitely repeated game.

A second function of charters was to set up and regulate the monitoring of inside and outside users of the fields. The monitoring system was organized by the community and involved designated guards who could impose fines on free-riders. The guards could not inflict physical punishment (which remained under state jurisdiction), and were incentivised by retaining a third of the fine. Reports of transgressions by community members were also incentivised in a similar way. Instead of letting the punishers bear the cost of monitoring, the *Carte* thus introduced mechanisms that alleviated the costs, and even made sanctioning a lucrative activity.

Nevertheless, the historical record reveals that fines were rarely imposed on insiders, but were mostly collected from trespassers (Casari 2007, p. 210). This could be because the rate of compliance was in fact very high inside each village, or because symbolic sanctions (like verbal reproach and gossip) were preferred when a member of the community was involved. Circulation of information and record-keeping were facilitated by holding regular meetings, with mandatory attendance for all community members. A special local court settled disputes among insiders and resolved ambiguous cases.

The case of Trentino's charters shows how the three main problems of folk-theorem mechanisms (infinite horizon, information, and costs) are solved by institutional design. Where these problems did not exist – or existed on a smaller scale – local villages were slower to adopt a charter, if they did adopt one at all. Smaller villages and communities in the most remote valleys of Trentino, for example, were less likely to adopt a charter than larger villages and communities in accessible and difficult-to-monitor locations (see Casari 2007, pp. 209–13).

The *Carte* are absolutely typical from this respect: Elinor Ostrom (1990), winner of the 2009 Nobel Memorial Prize, has identified the same features of Trentino's charters in a series of case studies spanning several centuries and countries across six continents. Stable membership, monitoring incentives, graduated fines, exclusion of outsiders, and conflict-resolution mechanisms figure in her list of key factors that make institutions for collective actions viable and robust across time. "In all known self-organized resource governance regimes that have survived for multiple generations, participants invest resources in monitoring and sanctioning the actions of each other so as to reduce the probability of free riding" (Ostrom 2000, p. 138). But the punishers are rewarded materially, and material damage is inflicted only rarely on the members

of the community. Most of the work is done by creating a long-term prospect for cooperation, and by the extensive use of symbolic sanctions.

Because Ostrom's work is sometimes cited by strong reciprocity theorists in support of their theses (see, e.g., Gintis et al. 2005), it is worth spending a few words on the implications of her work. Emphasis on the costs of punishment and the second-order dilemma these raise is indeed central in the common pool literature. The costs this literature refers to, however, are rather different from those modelled in strong reciprocity models of cooperation. Whereas Ostrom (1990) emphasises the cost of *setting up* common pool institutions, strong reciprocity theorists focus on the ongoing cost of *inflicting* punishment. These two problems are quite different and should be kept distinct.

Institutions such as the Trentino charters are in some respects more similar to national states in the way in which they administer sanctioning, than to the uncoordinated mechanisms of (most) punishment experiments. Once a coordinated punishment mechanism is in place, the cost of running it (and of implementing sanctions on a daily basis) largely takes care of itself. Common pool institutions avoid the problems caused by systems of uncoordinated punishment in which everyone decides on their own (arbitrarily and idiosyncratically) when and how to punish, with the potential for feuds that follows (Casari & Plott 2003). These advantages make coordinated punishment institutions remarkably robust and resilient across time.

Experiments performed by Yamagishi (1986) and Gürerk et al. (2006) have shown that subjects prefer and tend to migrate towards institutions with coordinated punishment; and a recent modelling exercise (Boyd et al. 2010) backs up the insight that these institutions may enjoy an evolutionary advantage compared to systems of uncoordinated sanctions. Boyd and co-workers propose a model in which the cost of punishment is inversely proportional to the number of punishers, and players can condition their decision on the size of the coalition. They show that under a plausible range of parameters cooperation and punishment can evolve. On the experimental side, Casari and Luini (2009) report higher levels of cooperation in Public Goods games when the decision to punish is supported by a coalition rather than by individual subjects. Part of the reason is that coordinated punishment tends to reduce individual attempts at anti-social punishment and revenge. Another advantage of real-life coordinated punishment is that it requires communication among peers, a factor that has a well-known positive effect on cooperation (Ostrom 1990; Ostrom et al. 1992). Communication in turn brings legitimacy – the punishment is perceived as just because it is consensual – and lack of legitimacy is probably a major cause of failure of externally imposed sanctions (Cardenas et al. 2000).

To sum up: Strong reciprocity theorists view punishment as *local*, *costly*, and *uncoordinated*. The empirical literature instead reports mainly the emergence of *local*, *cheap*, and *coordinated* punishment institutions. Both solutions to the dilemma of cooperation differ in part from the traditional imposition of external sanctions administered by the state, and both can be seen as raising second-order social dilemma problems. However, they also have rather different properties and should not be treated as if they were identical: The devil, in institutional design as in almost everything else, is very much in the details.

14. Models and policies

Having presented the bulk of the argument, I now turn to an obvious objection that can be raised against it. Lacking precise quantitative data, throughout this article, I have referred rather liberally to "cheap," "low-cost," and "no-cost" punishments as if they were the same thing. Undoubtedly, however, *a small cost is still a cost*, and for this reason alone, strong reciprocity theory can legitimately claim an advantage over its main rival.

This objection is far from trivial, and it raises important issues concerning the use of models and evidence in the social sciences. Part of my reluctance to speak of zero costs comes from the current lack of data concerning the cost-benefit ratio of punishment. And lacking precise data – on benefits especially – one should not rush to conclusions as soon as a small positive cost is detected. Nevertheless, I want to argue that even small but positive *net* costs would constitute too slender a basis to claim a victory for strong reciprocity theory.

The debate between weak and strong reciprocity theorists takes place in the context of an old controversy on the use of rational choice models – especially models based on narrow self-interest – in social policy. As Bowles and Gintis point out,

> Fehr and Gächter's (2002) experiment has implications for the design of constitutions and policies. It suggests that the objective should be to provide opportunities for the public-spirited to punish free riders, rather than to assume, as David Hume advised two-and-a-half centuries ago, that "every man ought to be supposed to be a knave and to have no other end, in all of his actions, than his private interest." (Bowles & Gintis 2002, pp. 127–28)

Using models to inform the design of institutions is a special activity that calls for special criteria of appraisal. The value of a policy-oriented model lies less in its descriptive accuracy than as a guide to effective *action*. This is particularly important in light of the well-known fact that all models simplify and betray reality in some respects. But while simplifications in one dimension ought to be exchanged for increased descriptive or predictive accuracy in some other dimension when we do pure science, simplifications ought to lead to *good advice* when policy-making is concerned.

How do weak and strong reciprocity fare in this respect? Costly punishment experiments are often accompanied – as in the preceding quotation – by suggestions that self-interest, long-term horizon, and information matter less than traditionally assumed by economists and biologists. But this suggestion is misleading. As Ostrom and others have emphasized, the opposite is likely to be true: Individual costs are crucial obstacles in the way to cooperation and must be kept low; uncoordinated punishment is dangerous and fragile; the shadow of the future and the circulation of information matter enormously. All these insights follow directly from weak reciprocity accounts of cooperation, in spite of the fact that its models – and their implications, like the folk theorem – are almost certainly false. False theories can still provide useful advice at the level of application.

Seen in this light, the issue of low- versus zero-cost punishment loses much of its importance. Perhaps gossip, ostracism, and verbal reproaches *are* a bit costly, and gene-culture co-evolution has helped humans overcome this little hurdle on the path towards sociality. Be that as it may, a theory of low-cost punishment would have relatively little practical interest for applied social science. Its advice for the policy-maker would be almost indistinguishable from that of weak reciprocity theory: Pay attention to individual costs; keep them low or make sure they are recouped later; extend the horizon of cooperation; and circulate information as much as possible. All these precepts were well known before the discovery of costly punishment in the laboratory, and the rise of strong reciprocity theory has only increased the risk that social scientists may forget about them.

15. Concluding remarks: Reciprocity without costly punishment

In this article I have argued the following:

1. Two interpretations of costly punishment experiments – narrow and wide – are usually conflated by strong reciprocity theorists.
2. Only the narrow interpretation is supported by experimental data, while the wide interpretation requires field evidence about the mechanisms that sustain cooperation in the wild.
3. Contrary to often-repeated claims, there is no evidence in the anthropological literature that costly material punishment is used in small acephalous societies, except in the regulation of sexual conflict.
4. On the contrary, there is a lot of evidence that revenge is a major cause of dissolution of social ties.
5. Economic cooperation in the small societies studied by anthropologists is usually supported by low-cost or no-cost mechanisms such as verbal criticism, ostracism, and coalitional punishment.
6. The robust common pools institutions studied by historians and institutional economists foster cooperation by recouping the costs of punishment, extending the horizon of cooperation, and circulating information among group members, as implied by weak reciprocity accounts of cooperation.

It is important to clarify that the evidence summarized in the preceding list does not refute the claim that Homo sapiens has evolved other-regarding ("social") preferences, or that punishment is an important mechanism for the enforcement of social norms. What it does challenge is the claim that social preferences are expressed via *costly* sanctions that sustain cooperation in a broad range of experimental and field situations. The weak point of strong reciprocity theory is not its analysis of individual motivation (Dubreuil 2010; Rosas 2008), but its narrow focus on artificial environments in which uncoordinated costly punishment has a beneficial effect on sociality. Strong reciprocity may well play a key motivating role in the creation of institutions – such as systems of collective monitoring and coordinated sanctioning – that foster cooperation, without triggering the negative side-effects of uncoordinated punishment.

In my view, the lack of support for the costly punishment account of cooperation is not to be celebrated. We would all like to have the best of both worlds: social cooperation in a large, diverse society without the burden of a centralized policing apparatus. But the evidence that cooperation can be sustained by decentralised costly punishment in the field is scant. Logically speaking, of course, we cannot rule out that in some cases costly punishment can sustain cooperation. However, while there is extensive evidence of spontaneously evolved institutions aimed at eliminating the cost of sanctioning, disregarding costs and relying on uncoordinated punishment would be very risky at the level of institutional design.

It is also worth stressing that lack of confirmation is not due to lack of scientificity. On the contrary, the rise of costly punishment is a good example of how the combination of rigorous theorizing with ingenious experimental data can foster quick progress in the social sciences. The moral to be drawn is that models and experiments can only take you so far, and the time has come for reciprocity theory to change gears and seek the test of historical and field data. This step was taken a long time ago in the investigation of related topics such as mutual insurance and collusion, and it is important to keep in mind that laboratory data – no matter how useful – cannot ultimately replace the evidence collected in the field.

Finally, nothing said in this article challenges the idea that strong *positive* reciprocity may be an important ingredient of human sociality. An adequate discussion of the other half of strong reciprocity would require a separate paper, but it will suffice to say that the prospects of positive reciprocity look brighter at first sight. Robust support comes from surveys (Andreoni et al. 1998; Fong 2001), laboratory (e.g., Berg et al. 1995; Burlando & Guala 2005; Fehr et al. 1993; Fischbacher et al. 2001), and natural field experiments (Frey & Meier 2004; Shang & Croson 2009).

This asymmetry of support is probably not an accident, and may reflect profound differences in the psychology of cost-processing. In the technical sense of economic theory, replying to a cooperative move with cooperation (instead of free-riding) in a one-shot dilemma game *is* equal to incurring a cost. Through the lens of the theory, positive reciprocity appears theoretically identical to negative reciprocity, for in both cases the agents are willing to pay a "fee" to reciprocate. But it is not obvious that positive and negative reciprocity are governed by the same psychological mechanisms. It is well known that the perception of gains and losses is biased by framing effects, and that missed opportunities are processed differently from directly incurred costs (e.g., Borges & Knetsch 1997; Kahneman et al. 1991). Psychological evidence on loss aversion suggests that we should be more reluctant to pay a fee to sanction nasty actions, than to miss an opportunity to profit at somebody else's expense. And it is possible that the evolutionarily ancient neural circuits that trigger negative reciprocity feelings work quite separately from the networks that support trust and positive reciprocity in the human brain (although the evidence is still contradictory and inconclusive; see, e.g., Tom et al. 2007; Yacubian et al. 2006).

Far from constituting an indictment of the strong reciprocity programme, then, the data call for a re-orientation away from its current obsession with costly punishment. More effort should be made in investigating how non-costly sanctions, backed up by adequate institutional

scaffoldings, may be used to sustain positive reciprocity in a variety of real-world settings (as in, e.g., Boyd et al. 2010; Rustagi et al. 2010). The policy implications of this insight are important enough to justify further investment in this research programme. But we should accept that accounts of cooperation based on costly, uncoordinated policing are not backed up by the empirical evidence collected so far.

ACKNOWLEDGMENTS

Previous versions of this article were presented at Bocconi University, the Max Planck Institute for Economics in Jena, and STOREP 2010. I would like to thank Paul Bloom, Ken Binmore, Paolo Garella, Herbert Gintis, Alessandro Innocenti, Josh Miller, Ivan Moscati, Elinor Ostrom, Nikos Nikiforakis, Alejandro Rosas, Don Ross, Polly Wiessner, and Jim Woodward for their help during revisions, as well as four anonymous referees for their generous comments. All the remaining mistakes are mine.

NOTES

1. Optimality in this article is intended in *material* terms, and the games (unless otherwise stated) are always specified in terms of material payoffs. This is a subtle but important point, since the behaviour of strong reciprocators can also be described as maximizing utility functions over non-material payoffs. Although the focus on material payoffs goes against the grain of some economic theorizing, it is more in line with evolutionary approaches where fitness usually tracks material gains.

2. There is a much older and prestigious research tradition in anthropology identifying reciprocity as a key force that keeps societies together (e.g., Gouldner 1960; Mauss 1954; Sahlins 1972/1974), but current theories rely almost exclusively on models and concepts introduced in the game theory and evolutionary biology literature of the 1970s.

3. There is evidence that the notion of "fair offer" varies across cultures. Although equal division in the Ultimatum Game is the modal offer in most Western societies, in Japan and Israel the mode goes down to 40% (Roth et al. 1991). Among the Au people of Papua New Guinea, the modal offer is in the region of 30%, and among the Hadza of East Africa, it is as low as 20%. The Machiguenga in Peru make the lowest offers observed in the Ultimatum Game so far (15%). Strong reciprocity theorists conclude that different norms of fair division can evolve in different contexts, and are supported by punishment mechanisms of the strong kind (Henrich et al. 2004).

4. I have used a non-standard matrix for presentational ease. The PP cell in Figure 2b is empty to indicate that sanctioned individuals typically do not have the option to counter-punish in these experiments. This is an important point, as we shall see later, for when counter-punishment is available, the experimental results change quite radically. Punishment of cooperators, corresponding to PC and CP, is usually possible and has led to interesting studies of antisocial punishment (e.g., Herrmann et al. 2008), but we shall ignore it for the time being.

5. Wilson (1979) calls behaviour of this kind "weakly altruistic" to highlight that it raises problems of selection, rather than motivation. The hypothesis put forward by Wilson is that weakly altruistic behaviour can evolve if the relative cost is low, because the force of group selection can compensate for the adverse effect of individual selection. Weak altruism is also the basis of Sober and Wilson's (1998) widely cited account of the evolution of moral norms.

6. Let us call b_i the benefit enjoyed by an individual i from consuming a public good produced by the individual's group. The *absolute* cost of punishment is c_i. Sanctions are costly in absolute terms if $c_i > b_i$. If $b_i > c_i$, then there is no absolute cost, and punishment is consistent with self-interest. The *relative* cost is the difference between the net benefit of the punisher $(b_i - c_i)$ and the benefit of the other group members who choose not to punish (b_j, for $j \neq i$). If $b_j > (b_i - c_i)$, there is a relative cost and individual selection works against the punisher. If $b_j = (b_i - c_i)$, there is no adverse selection within the group.

Open Peer Commentary

The social and psychological costs of punishing

doi:10.1017/S0140525X11001142

Gabrielle S. Adams and Elizabeth Mullen

Graduate School of Business, Stanford University, Stanford, CA 94305.

gsadams@stanford.edu **emullen@stanford.edu**

http://www.stanford.edu/~ gsadams

http://elizabeth.mullen.socialpsychology.org/

Abstract: We review evidence of the psychological and social costs associated with punishing. We propose that these psychological and social costs should be considered (in addition to material costs) when searching for evidence of costly punishment "in the wild."

In the target article, Guala argues that although costly punishment occurs in lab settings, there is little evidence of costly punishment "in the wild." Thus, he questions whether lab studies replicate the processes and conditions that support cooperation in the real world. We believe that Guala has defined "cost" too narrowly. Costs should include not only reduced material resources, but also decreased social status and psychological well-being. We argue that punishers often experience a number of social and psychological costs, and that such costs should be considered when searching for evidence of costly punishment "in the wild."

Punishers can experience social costs (e.g., reduced status), especially when observers question whether the punishment they enacted was proportionate to the seriousness of the transgression (Trevino 1992). For example, when employees observe managers punishing fellow employees in their organization, they sometimes report becoming less trusting, less respecting, and more fearful of those managers (Atwater et al. 2001). Feelings of distrust likely translate into less commitment to the manager (see Kramer & Cook [2004] for a review), which could lead to poorer outcomes for the punitive manager over time. Moreover, the fact that managers experience these social costs – even when punishing is somewhat legitimated by their social role – suggests that institutionalized punishment may not always mitigate the social costs of punishing. Research on the fundamental attribution error has demonstrated that people are prone to locating the cause of behavior within actors while failing to adequately account for the situational forces (e.g., roles) that may have caused the behavior (Ross 1977). As a result, people may still perceive punishers negatively (i.e., as aggressive or untrustworthy) even when punishment is encouraged by the punisher's role.

Enacting less harsh forms of punishment (such as ostracism, gossip, and verbal reproach) may also entail social or psychological costs. Although Guala argues that ostracism is not very costly, research suggests it is cognitively taxing and ego-depleting for the ostracizer. Participants who ostracized a confederate performed

worse on subsequent tests of their physical and mental capacities (e.g., they solved fewer anagrams) then participants who did not ostracize a confederate (Ciarocco et al. 2001). Even a seemingly innocuous form of punishment, gossip, is not without social costs: People who gossip negatively about others are less trusted and are more prone to negative reputations than those who do not gossip, even when controlling for the frequency of gossip (Turner et al. 2003). Finally, verbal reproach is also socially costly. Whistle-blowers who speak up against illegal behaviors perpetrated by employees of their organizations are susceptible to retaliation (e.g., negative performance evaluations, ostracism, dismissal) from members of the organization (Miceli et al. 2008; Near & Miceli 1995; Rothschild & Miethe 1999). Indeed, the prevalence of retaliation against whistle-blowers has led to the passage of legislation in the United States and other countries to attempt to protect whistle-blowers (see Miceli et al. [2008] for a review). Other examples of the costs of verbal reproach abound: Whites such as Viola Liuzzo who protested racial discrimination and segregation during the Civil Rights Movement in the United States suffered physical harm, reputational and material costs, and even death (Stanton 2000). In sum, ostracism, gossip, and verbal reproach can all be psychologically or socially costly forms of punishment. Although many factors likely influence whether these costs are experienced in any given situation, we simply highlight that punishers sometimes incur such costs.

Beyond these psychological and social costs, there is also anecdotal evidence of material costs associated with punishing "in the wild," such as when individuals or groups choose to boycott an organization. For example, the Dean and faculty at Vermont Law School denied military recruiters access to their campus facilities for many years because they opposed the military's "Don't Ask, Don't Tell" policy that prevents those who are openly gay, lesbian, or bisexual from serving in the military. As a result, the military had a difficult (but not impossible) time recruiting Vermont Law students, and the school forfeited approximately $500,000 in federal funding annually (Sanchez 2005).

Given the various types of costs we have reviewed, it is worth noting that empirical evidence supports Guala's speculation that people's emotions or motivations might lead them to punish even when it is against their immediate self-interest. Psychological research demonstrates that people's desires to punish are driven primarily by retribution, such that people punish to see the offenders suffer in a manner proportionate to their wrongdoing, even if the punishment will not effectively deter future transgressions (see Carlsmith & Darley [2008] for a review). In other words, people may punish to satisfy their retributive desires, even when it is costly to do so.

In conclusion, Guala dismisses non-material costs by claiming that they are not very costly or that they are not relevant to arguments of group fitness. In contrast, we argue that broadening the definition of costs to include social and psychological costs can help to inform the debate about whether there is evidence of costly punishment "in the wild."

Proximate and ultimate causes of punishment and strong reciprocity

doi:10.1017/S0140525X11001154

Pat Barclay

Department of Psychology, University of Guelph, East Guelph, Ontario, N1G 2W1, Canada.

barclayp@uoguelph.ca

http://www.uoguelph.ca/nacs/page.cfm?id = 229

Abstract: While admirable, Guala's discussion of reciprocity suffers from a confusion between proximate causes (psychological mechanisms triggering behaviour) and ultimate causes (evolved function of those psychological mechanisms). Because much work on "strong reciprocity" commits this error, I clarify the difference between proximate and ultimate causes of cooperation and punishment. I also caution against hasty rejections of "wide readings" of experimental evidence.

Guala reviews a number of interesting field studies that speak against the importance of punishment in maintaining cooperation. This is important because there is an abundance of laboratory research on punishment and cooperation which has outstripped the research in real-world settings. Underlying much of Guala's discussion of reciprocity and punishment, however, there lies confusion over proximate causation and ultimate causation. Confusion over these levels of analysis is not only present in Guala's target article, but is endemic to the entire field of cooperation and is particularly pronounced in the discussion of "strong reciprocity." This weakens Guala's arguments. In particular, it results in unwarranted statements against so-called weak reciprocity. As such, this topic requires clarification.

Any behaviour, including cooperation and punishment, can be explained at four different levels of analysis (Tinbergen 1968). Proximate causes include: (1) the psychological mechanisms that trigger behaviour (e.g., emotions, cognitions); and (2) the developmental processes that cause those psychological mechanisms to arise within an individual's lifetime (e.g., "innate" behaviour, learning, internalization of cultural norms). Ultimate causes include: (3) the evolutionary forces (e.g., reciprocity, mutualism, costly signalling) that result in those psychological mechanisms existing instead of other possible psychologies; and (4) the evolutionary history of those mechanisms and when they arose in our lineage (e.g., unique to humans, shared with other primates). These four levels of analysis – mechanism, development, function, and phylogeny – are complementary, not mutually exclusive. A complete explanation of any phenomenon requires an answer at each level.

An example can help clarify the proximate and ultimate causes of cooperative behaviour. Suppose that I genuinely value your welfare and I help you without any ulterior motives. If my action causes you to genuinely care about me, you will be more likely to help me when I need it, even when you anticipate no benefits for doing so. If I happen to find out, then your actions will cause me to value your welfare more and help you more often, and so on. The reciprocity in this example is not "weak": both of us unselfishly reciprocate "altruistic" acts. Both of us do benefit from helping each other, but neither one intended to benefit, and neither of us requires any foresight of the consequences. Helping can be altruistic from a proximate psychological perspective, but from an ultimate (functional) perspective it is advantageous to possess such a psychology. Thus, contrary to Guala's assumption, biologists do not assume *psychological* self-interest. To paraphrase Dawkins (1976/2006): The genes are selfish, but this doesn't mean the person is. One can make a similar argument with punishment: I may punish you because I am angry (proximate cause), and this may result in me receiving more future cooperation from you (potential ultimate cause of punitive sentiment), but this does not mean that my punishment was motivated by a desire for your cooperation.

Guala uses terms like "strong" and "weak" reciprocity, which are often misleading because they often conflate the proximate psychological mechanisms with the ultimate functional reasons for why those psychological mechanisms exist (Barclay 2010; West et al. 2007b). By itself, "strong reciprocity" is merely a description of behaviour, that is, the supposed tendency of people to cooperate, reward cooperators, and punish cooperators, even when there are no apparent benefits for doing so. The goal is to discover – at *all* levels among levels of analysis – why this tendency exists (if indeed it does). So-called theories of "weak reciprocity" are often theories about the ultimate function of cooperative and punitive sentiment, not theories about what specifically that sentiment is or how it develops. People possess certain emotions and psychological mechanisms which

are predicted to be adaptive *on average* outside the laboratory; for example, if being nice invites reciprocation. People bring these psychological mechanisms with them into the laboratory, where the behaviour produced may or may not still be adaptive on average (Barclay, 2011; West et al. 2011). "Maladaptive" behaviour can persist despite repeated anonymous encounters, as long as the same proximate psychological mechanisms are repeatedly triggered (e.g., anger, desire for fairness, empathy). However, this would say little about the ultimate function that those mechanisms serve outside the laboratory. Too much ink has been spilled by researchers who do not realize that their colleagues are simply addressing a different level of analysis.

On a completely different note, Guala makes a useful distinction between wide and narrow readings of the experimental evidence, and what each reading implies. Wide interpretations can clearly be taken too far: If punishment (or any other phenomenon) supports cooperation in the lab, it does not necessarily mean that this is what supports it outside the lab. However, I would caution against hasty abandonment of such wide interpretations. Sometimes laboratory experiments use controlled conditions to test whether a proposed mechanism *could* support punishment. At other times, such experiments test the validity of theories of human behaviour (Mook 1983): If a predicted phenomenon cannot be found in the lab under ideal controlled conditions, then we must either reject or revise any theory that relies on that phenomenon (see, e.g., the lack of punishment towards non-punishers in Kiyonari & Barclay 2008). If successful, do these findings need confirmatory non-laboratory observations with real-world phenomena? Absolutely. Convergent evidence is crucial in all scientific enterprises, and the laboratory and the field have their own respective strengths and weaknesses. As such, we should all strongly support the call for collaborations across disciplines and between the lab and the field. Guala's target article has clearly shown that the punishment literature needs more of this, and for that it should be commended.

ACKNOWLEDGMENTS

I thank Daniel Krupp for discussions and the Social Sciences and Humanities Research Council of Canada (SSHRC) for support.

The restorative logic of punishment: Another argument in favor of weak selection

doi:10.1017/S0140525X11001166

Nicolas Baumard

Philosophy, Politics, and Economics Program, University of Pennsylvania, Philadelphia, PA 19104.

nbaumard@gmail.com

https://sites.google.com/site/nicolasbaumard/Home

Abstract: Strong reciprocity theorists claim that punishment has evolved to promote the good of the group and to deter cheating. By contrast, weak reciprocity suggests that punishment aims to restore justice (i.e., reciprocity) between the criminal and his victim. Experimental evidences as well as field observations suggest that humans punish criminals to restore fairness rather than to support group cooperation.

As Guala rightly notes, there is very little evidence that punishment plays a role in the stabilization of cooperation in small-scale societies. On the other hand, as he also notes, it is difficult to totally rule out the strong view of punishment as it is complicated to precisely assess the costs of punishment in the field (Are there really no costs in punishing others? Aren't there many hidden benefits for the individual who punish? etc.). There is, however, another way to disentangle the two views of punishment, namely, the forms that punishments take. Indeed, the two theories – the weak and the strong – make different predictions regarding the logic of punishment.

Group selection theory holds that punishment aims to promote the good of the group by sustaining cooperation and preventing cheating (Boyd et al. 2003; Fehr & Gächter 2002; Henrich & Boyd 2001). This implies that punishment should be calibrated to deter crimes and render them non-advantageous. Here, group selection parallels the utilitarian doctrine of punishment, which contends that punishment should be used to deter crimes and maximize the good of society (Polinsky & Shavell 2000; Posner 1983). The utilitarian theory of punishment holds, for instance, that the detection rate of a given crime and the publicity associated with a given conviction are relevant factors in assigning punishments. If a crime is difficult to detect, the punishment for that crime ought to be made more severe in order to counterbalance the temptation created by the low risk of getting caught. Likewise, if a conviction is likely to get a lot of publicity, a law enforcement system interested in deterrence should take advantage of this circumstance by "making an example" of the convict with a particularly severe punishment, thus getting a maximum of deterrence for its punishment.

By contrast, individual selection predicts a "restorative" or "retributive" logic for punishment (Baumard 2011). Restorative logic holds that punishment aims to restore justice between the criminal and the victim – either by harming the criminal or by compensating the victim. In intuitive terms, people are punished because they "deserve" to be punished, and not because punishing them would be useful for the society at large.

This restorative logic is a direct consequence of the way cooperation has evolved among humans (Baumard 2010a; Trivers 1971). Indeed, human beings belong to a highly cooperative species and get most of their resources from collective actions, solidarity, exchanges, and so forth. (Gurven 2004; Hill & Kaplan 1999). In the ancestral environment, individuals were in competition to be recruited for the most fruitful ventures, and it was vital to share the benefits of cooperation in a mutually advantageous manner. If individuals took a bigger share of the benefits, their partners would leave them for more interesting partners. If they took a smaller share, they would be exploited by their partners who would receive more than what they had contributed to produce. This competition to attract cooperative partners is thus likely to have led to selection for a "sense of fairness," a cognitive device that motivates individuals to share the costs and benefits of social interaction in an impartial way (André & Baumard 2011). If cooperation is based on fairness, then crimes create an unfair relationship between the criminal and her victim, and people have the intuition that the criminal ought to compensate the victim or to be punished in order to restore justice.

It is worth mentioning that this theory does not mean that punishment should be absent in human societies. As Guala notes, modern societies have found many institutional ways to reduce the costs of punishments. Although these institutions are absent in smaller societies, justice can still be restored by individuals seeking to retaliate. Retaliation is indeed advantageous from an individual perspective and can indeed be found in many nonhuman species (Clutton-Brock & Parker 1995). As Evans-Pritchard noted, in societies where there is no penal system, "self-help, with some backing of public opinion, is the main sanction" (Evans-Pritchard 1940/1969, p. 169).

In this kind of situations, selfish and moral motives converge: The victim (or his allies) attacks the criminal to signal his strength and gains a reputation as someone who cannot be attacked without risk; and by doing so, he also punishes the wrongdoer by allowing justice to be done. In line with this idea, people in small-scale societies distinguish between legitimate (and proportionate) retaliation and illegitimate (and disproportionate) retaliation (von Fürer-Hameindorf 1967; Miller 1990). Retaliation is thus clearly limited by moral concerns: within the group, it has to be proportionate to the prejudice. As the *Lex Talionis* says, "an eye for an eye, a tooth for a tooth," but no more.

Individual selection thus clearly predicts some kind of punishment, and, more importantly, it predicts that punishments should aim toward a specific goal (restoring fairness) that differs from the utilitarian goal predicted by group selection (preventing wrongdoing). Experimental studies, relying on a variety of methodologies, suggest that punishments fit individual selection more than group selection. Indeed, when people punish harmdoers, they generally respond to factors relevant to a retributive theory of punishment (magnitude of harm, moral intentions) and ignore factors relevant to the group selection theory (likelihood of detection, publicity, likelihood of repeat offending) (Baron et al. 1993; Baron & Ritov 2008; Carlsmith et al. 2002; Darley et al. 2000; Glaeser & Sacerdote 2000; Sunstein et al. 2000).

In line with these results, field observations have extensively demonstrated that, in keeping with the prediction, the level of compensation in stateless societies is directly proportional to the prejudice inflicted to the victim: For example, the wrongdoer owes more to the victim if the wrongdoer has killed a family member or eloped with a wife than if he has stolen animals or destroyed crops (Hoebel 1954; Howell 1954; Malinowski 1926). To conclude, punishment does not seem to be a group adaptation. It follows the logic of fairness rather than the interests of the group.

Reciprocity and uncertainty

doi:10.1017/S0140525X11001178

Yoella Bereby-Meyer

Department of Psychology, Ben-Gurion University of the Negev, Beer Sheva 84105, Israel.

Yoella@bgu.ac.il www.YoellaBerebyMeyer.com

Abstract: Guala points to a discrepancy between strong negative reciprocity observed in the lab and the way cooperation is sustained "in the wild." This commentary suggests that in lab experiments, strong negative reciprocity is limited when uncertainty exists regarding the players' actions and the intentions. Thus, costly punishment is indeed a limited mechanism for sustaining cooperation in an uncertain environment.

Strong reciprocity is the behavioral predisposition to cooperate conditionally on others' cooperation and to punish violations of cooperative norms even at a net cost to the punisher (Fehr & Gintis 2007). The phenomenon has been the subject of considerable research in the last few decades (e.g., Camerer 2003; Fehr & Gächter 2000b; Rabin 1993), and its existence is well established.

In the target article, Guala points to a discrepancy between the strong negative reciprocity that is observed in the lab and the way cooperation is sustained "in the wild." Specifically, he suggests that there is no indication for costly punishment in the wild. This claim gives rise to the question as to what extent one can predict actual behavior in real-life situations from behavior in the very artificial and contrived laboratory setting. The author suggests that behavior in the laboratory with respect to strong negative reciprocity does not really reflect behavior in real life. However, the matter may actually be somewhat more complex than it seems. Even if there is no strong negative reciprocity in the real world, this may still be in line with the results from laboratory studies. One simply has to make sure that the laboratory studies capture crucial characteristics of the real world.

One of the main properties of real-world situations is some degree of uncertainty (which does not usually exist in laboratory studies and particularly in those the author referred to). In many real-life social dilemmas people face uncertainty of two types: (1) environmental uncertainty, which is uncertainty regarding aspects of the dilemma (e.g., the size of the common resource); and (2) social uncertainty, which is uncertainty regarding the other group members' choices (Messick et al. 1998). Moreover, outcomes may be determined probabilistically.

For negative reciprocity to occur, accurate knowledge regarding the actions and the intentions of the players is important. If uncertainty exists, it will be difficult to determine whether the action and the outcome were the result of violations of cooperative norms. While most laboratory experiments have dealt with situations that are certain, a number of studies have introduced some degree of uncertainty into situations in which negative reciprocity is possible. These studies consistently show that uncertainty lowers the tendency towards negative reciprocity.

Most of the evidence for strong negative reciprocity was observed in the Ultimatum Game (UG). In research on the UG, responders are very likely to reject offers that are less than 30% of the cake (e.g., Güth et al. 1982). By rejecting the offer, the responder gives up a possible gain; and thus the finding is interpreted as evidence for responders' willingness to pay a cost to punish the proposers they perceive as acting unfairly, even if they will never meet the proposers again. In the classic experiment and in most following ones, the size of the pie to be divided is common knowledge. As was noted by Croson (1996), this assumption is unrealistic.

Several experiments investigated UGs with one-sided uncertainty on the part of the responder. Typically, in these experiments proposers know the exact amount of money to be divided, and responders either know nothing at all or know the probability distribution of possible amounts.

In most studies responders accepted lower offers when they did not know the size of the pie and when the lack of information was common knowledge. Proposers, in turn, did not hesitate to exploit this behavior and offered little when the amount to divide was large (e.g., Croson 1996; Mitzkewitz & Nagel 1993; Rapoport et al. 1996). Recently Gehrig et al. (2007) studied a UG with a different source of uncertainty. In their game the responder knows the pie size but not the offer when deciding whether to accept or reject (i.e., has imperfect information). Responders never reject in this game, even when they anticipate low offers.

Under both types of uncertainty responders seem to give proposers the benefit of the doubt: Because a low offer could be fair if the pie is small or the yet-unknown offer could eventually be fair, rejecting the offer would mean punishing the proposer unfairly. Consequently, with uncertainty lower offers are more likely to be accepted. This behavior is strong evidence that rejections in the UG are an expression of preference when responders do know the proposer's payoff (Camerer 2003); and therefore the ability to generalize these preferences to situations with uncertainty is limited.

Uncertainty also affects reciprocity in repeated interactions. The ability of reciprocity to sustain cooperation in the long run, and specifically in the iterated Prisoner's Dilemma, was demonstrated by Axelrod's (1984) well-known computer tournaments. Later it has been shown that cooperation is much more difficult to maintain if there is uncertainty regarding players' actions, that is, if there is random error either in choosing actions or in monitoring others' actions (see, e.g., Axelrod & Dion 1988; Bendor 1993; Green & Porter 1984; Sainty 1999). That is, if actions are noisy, a player does not know whether another player's defection was an error or an intended choice, and strategies involving reciprocation (e.g., tit for tat) can break down. But even if players can monitor others' past actions perfectly in a repeated Prisoner's Dilemma game, if payoffs are noisy, players learn to cooperate much less (e.g., Bereby-Meyer & Roth 2006; Kunreuther et al. 2009).

Hence, the fact that Guala in his analysis of real-world situations did not find evidence for strong negative reciprocity does not necessarily imply that results from laboratory studies cannot predict reciprocity behavior in the real world. Instead, one can perhaps conclude that in situations with uncertainty,

costly punishment is less likely to occur, and therefore in these situations probably other mechanisms are needed to sustain cooperation.

Costs and benefits in hunter-gatherer punishment

doi:10.1017/S0140525X11001403

Christopher Boehm

Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089.

cboehm1@msn.com

Abstract: Hunter-gatherer punishment involves costs and benefits to individuals and groups, but the costs do not necessarily fit with the assumptions made in models that consider punishment to be altruistic – which brings in the free-rider problem and the problem of second-order free-riders. In this commentary, I present foragers' capital punishment patterns ethnographically, in the interest of establishing whether such punishment is likely to be costly; and I suggest that in many cases abstentions from punishment that might be taken as defections by free-riders are actually caused by social-structural considerations rather than being an effect of free-rider genes. This presentation of data supplements the ethnographic analysis provided by Guala.

If one is interested in explaining both the social dynamics and the genetics of human punishment, the everyday behaviors of Late-Pleistocene type foragers are of special interest, even though their ethnographic description is neither complete nor even consistent. There exist some remarkable and strong social central tendencies among the 150-plus documented foraging societies that qualify as "Late-Pleistocene appropriate" (LPA), for these people cluster in mobile egalitarian bands, form moral communities, condemn and punish predatory behaviors like bullying and cheating, and actively favor altruistic cooperation (Boehm 2000; 2008). Two important social goals are: (1) keeping political life egalitarian and (2) promotion of cooperation; and their methods range from shaming, to ostracism, to capital punishment.

In terms of assessing punishment's costs and possible second-order free-rider problems, capital punishment is of special interest because deviants are likely to resist being killed; furthermore, they may be avenged. My coded data on 50 LPA societies (see Boehm, in press) reveal patterns that complement Guala's analysis (see Table 1).

At present there is an Inuit and Australian Aborigine bias in this sample, which may be skewing the data somewhat in favor of sorcery. More generally, the data will be inherently "incomplete" due to reticence, because colonial administrations punish indigenously legitimate executions as murder (see Lee 1979); but, with that caveat, the main targeted deviant pattern involves forceful personalities that go against the egalitarian grain, while (with much smaller numbers) devious predators and sexual malefactors seem to come in second. I would suggest that over the millennia all LPA foragers have been executing serious deviants (see also, Otterbein 1986) on a rare but significant basis, and that the main culprits have been would-be dominators such as sorcerers or serial killers.

Mobile foragers live in groups averaging 20–25 persons, which are composed largely of nonrelatives or distant relatives (Hill et al. 2011) but often contain pairs of siblings. There is sometimes the possibility that a male may avenge a close kinsman's death even if the victim was a major social deviant (e.g., Boehm 2011; van den Steenhoven 1962), so potentially the risks in using capital punishment were high unless the problems of predictable resistance and possible retaliation could be coped with.

In another analysis that concentrates on LPA foragers' methods of social control, in a smaller sample of 10 societies I discovered that in six of them capital punishment was done by delegating a close kinsman to kill the culprit by ambush (see also Woodburn 1982), while in six (mainly overlapping) societies in the same sample ethnographers reported merely that "the group" killed the culprit.

Often it is impossible to tell whether these group actions involved collective killing or delegation, but out of 22 Bushman homicide cases there is one well-described account of an assassination attempt that turns collective; first a man hits a serial killer with poisoned arrows when he is awake, then the latter wounds a woman and kills her husband, and finally the entire group attacks him as he is dying of the poison (Lee 1979). The impression is of a group agreement that the man must go, but of great inefficiency. In a *tribal* example of a highly efficient communal execution, it is clear that the point of everybody participating is to avoid retaliation by making it impossible to determine who actually killed the culprit (Boehm 1986).

Risks are greatly reduced when a band delegates a close male kinsman of the target to do the job: first, because he will ambush the deviant in his sleep; and second, because otherwise predictable lethal retaliation will be set aside because of the kin tie. But a remaining evolutionary paradox is that the delegated executioner faces costs (the slight risk of the target fighting back, and the definite loss of a close male kinsman), while the

Table 1 (Boehm). *Capital Punishment in 50 LPA Foraging Societies*

Type of Deviance	Specific Deviances	Societies Reporting
Intimidates Group	Intimidation through malicious sorcery	11
	Repeated murder	5
	Action otherwise as tyrant	3
	Psychotic aggression	2
Cunning Deviance	Theft	1
	Cheating (meat-sharing context)	1
Sexual Transgression	Incest	3
	Adultery	2
	Premarital sex	1
Miscellaneous	Violation of taboo (endangering the group)	5
	Betrayal of group to outsiders	2
	"Serious" or "Shocking" transgression	2
Deviance Unspecified		7
Total Societies Reporting Capital Punishment		24

rest of the band (non-kin and distant kin) can be seen as free-riders who benefit substantially but pay no costs.

However, there is another way setting up the genetic cost/benefit analysis. The executioner who pays such costs is merely caught in a structural position, in which he becomes the chosen executioner because he is close kin, whereas the free-rider roles of those who abstain are also determined by social position. Thus, free-rider *genes* are not at issue because the free-riding is determined *situationally*.

In this light, we may reconsider the ethnographically well-described Mbuti case Guala cites from Turnbull (1961). Cephu cheats on a meat-acquisition system which is designed to bring in a fair share of game for all the participating families; and, collectively, most of the band actively shames him in ways that are humiliating while Cephu's loyal followers stand aside – but do not actively back him. This too is situational, because they are kin. It is worth noting that the sanctioning goes beyond shaming when one band member threatens the arrogant Cephu with ejection from the band; but he is taking little risk because the people backing him are in a state of moral outrage.

I emphasize that the several families associated closely with Cephu likely would be conventionally modeled as free-riding defectors because they stand aside; and also, that in fact this is not a matter of opportunistic free-rider genes in action. It is simply a situational matter, and over the millennia such stepping aside has had nothing to do with genes. In such contexts, the free-rider problem does not apply.

Guala has opened up some interesting questions, and has used ethnographic data in doing so. Perhaps these further ethnographic nuances may serve as useful food for thought, for scholars who use experiments with students (or non-LPA nonliterates) to try to understand human nature.

ACKNOWLEDGMENT

I thank Joe Henrich for comments on this commentary.

The punishment that sustains cooperation is often coordinated and costly

doi:10.1017/S0140525X1100118X

Samuel Bowles,[a] Robert Boyd,[b] Sarah Mathew,[b] and Peter J. Richerson[c]

[a]*Santa Fe Institute, Santa Fe, NM 87501;* [b]*Department of Anthropology, University of California–Los Angeles, Los Angeles, CA 90095;* [c]*Department of Environmental Science and Policy, University of California–Davis, Davis CA 95616.*

samuel.bowles@gmail.com **rboyd@anthro.ucla.edu**
smathew@ucla.edu **pjricherson@ucdavis.edu**
http://www.santafe.edu/~bowles
http://www.sscnet.ucla.edu/anthro/faculty/boyd/
http://smathew.bol.ucla.edu/Site/Home.html
http://www.des.ucdavis.edu/faculty/Richerson/Richerson.htm

Abstract: Experiments are not models of cooperation; instead, they demonstrate the presence of the ethical and other-regarding predispositions that often motivate cooperation and the punishment of free-riders. Experimental behavior predicts subjects' cooperation in the field. Ethnographic studies in small-scale societies without formal coercive institutions demonstrate that disciplining defectors is both essential to cooperation and often costly to the punisher.

We are grateful to Francesco Guala for providing a thoughtful reflection on what recent social dilemma experiments can tell us about real-world cooperation and the need for complementary ethnographic, historical approaches. But Guala's contribution is packaged along with what we think is a misunderstanding of our work, an overly pessimistic appraisal of the external validity of experimental results, and a very partial reading of the evidence on costly punishment in small-scale societies.

The core of strong reciprocity is that human cooperation cannot be understood entirely as the result of repeated social interaction and self-interested individual calculation. Instead, people are motivated to cooperate with one another and to punish free-riding by a variety of ethical and other-regarding motives. Guala gets this right. However, he incorrectly believes that strong reciprocity requires punishment to be both very costly and uncoordinated. Punishment is costly when the cost of administering punishment, however small, exceeds the private benefit it creates for the punisher, thus giving rise to a second-order free-rider problem. Mechanisms like conformism, kin selection, or cultural group selection can solve the second-order free-rider problem, but usually only if the cost of punishment is low, either because it is rare (e.g., Boyd et al. 2003; Henrich & Boyd 2001) or because it is collectively administered (Boyd et al. 2010).

Everyday social life, even among strangers, is regulated by many individual acts of uncoordinated punishment. We are all aware of the pain we experience when we are frowned upon in public places among strangers. However, we agree with Guala that more costly forms of punishment in natural settings are usually collective. We capture this in our paper "Coordinated Punishment of Defectors Sustains Cooperation…" (Boyd et al. 2010), which Guala cites but seems to have misunderstood. In this model, potential punishers signal their willingness to punish, but they punish free-riders only when enough fellow punishers have signaled. When there is no assortment, there are two possible evolutionary equilibria: a population without punishment or cooperation, and a population with a mix of punishers and non-punishers in which most actors cooperate. Mean fitness is higher when punishers are present. When we allow an empirically realistic degree of assortment in the population, punishment may proliferate even when rare; and when it does, it is altruistic.

We developed this model because we share Guala's dissatisfaction with the typical representation of punishment as an individual act rather than something deliberated on by groups and undertaken jointly (but see Ertan et al. 2009). Nonetheless, experiments make a major contribution by showing that the predispositions that motivate punishment are common in many populations. We agree with Guala that we need better tests of the external validity of these experimental results. But two kinds of evidence are encouraging.

First, behavior in experiments predicts subjects' cooperation in the field. Brazilian shrimpers use large plastic bucket-like contraptions in which holes are cut to allow the immature shrimp to escape, thereby preserving the stock for future catches. Because they can cut holes of any size, the fishermen face a real-world social dilemma. Large holes represent cooperation with other fishers; small trap holes are a form of defection, and – just as in the Public Goods Game – having small holes is the dominant strategy for a self-interested shrimper. Not surprisingly, those who contributed most in a public goods experiment were also those who cut larger holes in their traps (Fehr & Leibbrandt 2011). The effects, controlling for a number of other possible influences on hole size, are substantial.

Additional evidence of external validity comes from a set of experiments and field studies with 49 groups of herders of the Bale Oromo people in Ethiopia, who were engaged in forest commons management (Rustagi et al. 2010, which Guala cites). The most common behavioral type in the experiments, constituting a bit more than a third of the subjects, were "conditional cooperators" who responded positively to higher contributions by others. Controlling for a large number of other influences on the success of the forest projects, the authors found that groups with more conditional cooperators planted more trees. (See Bowles & Gintis [2011] for more evidence on external validity.)

Second, there is much evidence for costly third-party punishment among societies without formal coercive institutions. Mathew and Boyd (2011) present extensive quantitative data showing that punishment of cowardice and other forms of free-riding plays an important role in warfare among the Turkana, an acephalous African pastoral group. Community members decide whether a violation occurred, and if it has, corporal punishment is administered by the violator's age-mates, not those damaged by the violation. Punishing takes time and effort and may damage valuable social relationships.

Contrary to Guala, punishment has been observed in the simplest foraging societies. Among the Walbiri of Australia, for example, offenses like homicide, physical assault, sacrilege, adultery, and theft were punished by death, wounding with a spear or knife, or attack with a club or boomerang (Meggitt 1962, pp. 256–59). The local community determined whether the act was an offense, decided on the punishment, nominated the person to carry out the punishment, and appointed the people responsible for ensuring that the punisher does not face retaliation (p. 255).

In some cases, meting out punishment is very costly. Among Aranda foragers of the Central Desert in Australia, wrongdoers were sometimes executed. The elders collectively decided on the fate of the wrongdoer, and assigned a group of young men to carry out the execution. Strehlow (1970, pp. 117–18) describes two cases in which the violator's relatives did not think the execution was justified, and killed the young men who had carried it out. According to Strehlow, capital punishment of this nature occurred in all Central Australian tribes before colonial administration made them a criminal offense.

Weak reciprocity alone cannot explain peer punishment

doi:10.1017/S0140525X11001191

Marco Casari

Department of Economics, University of Bologna, Piazza Scaravilli 2, 40126 Bologna, Italy.

marco.casari@unibo.it http://www2.dse.unibo.it/casari/

Abstract: The claims about (1) the lack of empirical support for a model of strong reciprocation and (2) the irrelevant empirical role of costly punishment to support cooperation in the field need qualifications. The interpretation of field evidence is not straightforward, and other-regarding preferences are also likely to play a role in the field.

Guala should be praised for having raised this debate about punishment experiments. I will focus on two main points. First, the target article claims that the empirical evidence on peer punishment is not enough to support theories based on strong reciprocity. As I argue below, behavior in peer punishment experiments cannot be entirely rationalized with self-regarding or weak reciprocity attitudes, and strong reciprocity is one model of other-regarding behavior among others currently under debate.

There is no lack of anecdotes about peer pressure and punishment in field settings, ranging from high school students to miners on strike (Francis 1985) to fishermen communities (Bromley 1992) to workplaces (Kandel & Lazear 1992). In the region studied in Casari (2007), costly punishment is still practised today. Recently, 1,800 young grapevines have been cut with pruning hooks and shears. Apparently two people acted overnight, causing damage in thousands of euros. In the last five years, there have been seven similar episodes in the same community. Generally the culprits remain unknown (Nardon 2011). The issue of peer punishment was raised after field research and was not born as a laboratory anomaly. Experiments helped to clarify the extent and drivers of peer punishment, because field evidence is often hard to interpret. There are nuisance factors and measurement limitations: The interaction may be repeated, the fine-to-fee ratio unknown, or institutions to promote cooperation may be present. Controlled experiments are useful because they remove many of these limitations. One robust finding is the willingness of many people to pay a personal cost to inflict a punishment on others, especially on free-riders. This result persists in one-shot situations when the punisher incurs a material loss. As in other experiments, the data point toward the existence of a mix of motivations in economic decision making. While most subjects exhibit exclusively self-regarding motivations, there are others who also exhibit an array of other-regarding motivations.

Weak reciprocity is simply not enough to rationalize the existing experimental results on peer punishment. For instance, subjects do not treat peer punishment as a second-order public good, that is, they do not employ punishment mainly to provide incentives for the free-rider to contribute, as a weak reciprocity argument would suggest (Casari & Luini 2006; 2009). One can also experiment settings with indefinite repetition, where weak reciprocators can support cooperative outcomes through a rational strategy different than costly peer punishment. When four subjects indefinitely played prisoner's dilemmas in random pairs, more than half of the time cooperators targeted defectors with peer punishment (Camera & Casari 2009). Rational, self-regarding subjects had the alternative to support full cooperation through a simple grim trigger strategy. Instead, many still employed peer punishment. To sum up, experiments on peer punishment have shed light over important aspects of cooperative behavior that are likely to apply also in field situations. Yet, the existing evidence still leaves some deep questions open about the genetic versus cultural origin of other-regarding motivations; about the degree of external validity of experiments; and, about what model can fit the observed patterns of punishment with reasonable precision.

Guala's second main point is that cooperation in the field does not rely primarily on the forces uncovered in punishment experiments but is promoted by institutions that reduce the costs of decentralized punishment and facilitate the functioning of weak reciprocity mechanisms. I agree, although I will discuss two of Guala's related statements, which are based on unconvincing interpretations of the anthropological evidence: (1) Peer punishment does not occur in the field; (2) hence, it is irrelevant in a field setting. Guala argues that peer punishment is rarely employed and that some punishment acts are not costly, given that the cost to inflict punishment is claimed to be "low." In the literature, what matters is the fine-to-fee ratio of a punishment act, not simply the absolute cost of a punishment request. Moreover, sanctions ought not to be always large but, rather, graduated (Ostrom 1990). In the lab one observes a proportion between crime and punishment, that is, actions of full free-riding attract more punishment than actions of partial free-riding, and something similar may be expected in the field.

When extrapolating to field situations, one has to keep in mind that in laboratory experiments, people are forced to interact with others, have little control over the information flow, and have only few options available. In the field, people have multiple ways to inflict punishment and have strategies alternative to peer punishment. Instead of physically confronting a norm violator, a cooperator may decide to act to lower the cost to punish, to create institutions, or to move camp elsewhere. Hence, people can optimize over the many strategies available. A lower-than-expected frequency of peer punishment actions may simply reveal that there are better strategies in that situation, not that they are unavailable or irrelevant. For instance, speaking up against someone is costly because it exposes one to the risk of retaliation (Wiessner 2005), as whistle-blowers know. To avoid counter-punishment, in the field people may increase the level of anonymity by spreading gossip instead of reproaching

someone face-to-face. Another way to dilute the risk of retaliation is to punish in coalition with others, which is another form of decentralized punishment (Casari & Luini 2009).

The success in overcoming a social dilemma situation may be due to multiple factors, and other-regarding attitudes may be one of them. Although in the field they can be sometimes hard to quantify, they may nevertheless play an important role. The analysis of the *Carte di Regola* followed a canonical model with identical, self-regarding agents out of parsimony in organizing the historical evidence (Casari 2007). A related experiment examined in more depth a specific feature of the *Carte* system and uncovered the subtle role of subjects' heterogeneity in behavior for the success of the institution. Under a *Carte* system, the interaction of pro-social, self-regarding, and anti-social attitudes increased group welfare (Casari & Plott 2003). Field and experimental evidence, therefore, complement each other.

In medio stat virtus: Theoretical and methodological extremes regarding reciprocity will not explain complex social behaviors

doi:10.1017/S0140525X11001208

Claudia Civai and Alan Langus

SISSA/ISAS–International School for Advanced Studies, via Bonomea, 265, 34136 Trieste, Italy.

civai@sissa.it alanlangus@gmail.com

http://www.sissa.it/cns/index.html

Abstract: Guala contests the validity of strong reciprocity as a key element in shaping social behavior by contrasting evidence from experimental games to that of natural and historic data. He suggests that in order to understand the evolution of social behavior researchers should focus on natural data and weak reciprocity. We disagree with Guala's proposal to shift the focus of the study from one extreme of the spectrum (strong reciprocity) to the other extreme (weak reciprocity). We argue that the study of the evolution of social behavior must be comparative in nature, and we point out experimental evidence that shows that social behavior is not cooperation determined by a set of fixed factors. We argue for a model that sees social behavior as a dynamic interaction of genetic and environmental factors.

The target article discusses reciprocity in human social behavior. Guala argues that the evidence for strong reciprocity (i.e., high-cost mechanisms of punishment) found in experimental games such as the Ultimatum Game does not coincide with the evidence for weak reciprocity (i.e., low-cost or no-cost mechanisms) found in anthropological and historical data. Guala suggests that one possible solution to this problem may lie in the fact that weak reciprocity with low-cost or no-cost mechanisms is more relevant in natural situations. According to Guala, the study of social behavior should therefore focus on field experiments and analysis of historical as well as anthropological data, rather than controlled laboratory experiments.

Contrary to Guala's belief, this methodological shift is unlikely to add explanatory power to models of reciprocity in evolutionary scale. Comparative studies suggest that reciprocity exists in nonhuman animals (Brosnan & de Waal 2003; Dufour et al. 2009; Fruteau et al. 2009) and that even monkeys use a combination of high-cost and low-cost mechanisms to endorse cooperative behavior – for example, physical fights between males to establish the group leader and ignorance by group members leading the losing male to leave the group. The evolution of behavior, which depends on the interaction of genetic mechanisms and environmental factors, cannot be captured by historic data from the Middle Ages or by anthropological evidence that assimilates behavioral norms of small societies to those of evolutionarily older communities. In fact, even though genes can be regulated according to environmental factors (Robinson et al. 2008), the molecular mechanisms capable of shaping complex behavior appear to be very conservative across species (Krieger & Ross 2002; Langus et al., in press). This means that the fundamental mechanisms shaping social behavior are likely to be shared with other nonhuman animals. Guala's argument that strong reciprocity is absent even in primitive communities thus suggests that there have been no evolutionary changes in reciprocity within the human lineage. Shifting between experimental approaches (laboratory vs. natural settings) and theoretical extremes (strong vs. weak reciprocity) is therefore unlikely to be sufficient to capture the fine-tuned fabric and the evolution of human social behavior.

We believe that the problem with models of reciprocity is neither the methodological approach, nor whether one chooses to believe in strong or weak reciprocity. We argue that a unique theory that defines social behaviors as cooperation and presupposes the existence of a standard rational behavior or a standard optimal strategy (e.g., Dufwenberg & Kirchsteiger 2004; Fehr & Schmidt 1999) is essentially flawed. In humans, social behavior does not solely depend on fixed factors triggering automatic mechanisms (e.g., emotions, internalized norms, or social preferences). For example, negative emotions have been considered to be the ultimate cause that explains "irrational" reactions to unfairness (Pillutla et al. 1996; Sanfey et al. 2003; van't Wout et al. 2006). However, Civai et al. (2010a) show that emotions are not necessarily correlated to cooperation. In a modified version of the Ultimatum Game, when participants were directly involved in the bargaining process, their skin-conductance response correlated with rejection of unfairness, whereas when they had to decide on behalf of a third party, their electrodermic response did not show a significant correlation with unfairness. This evidence is supported by fMRI findings that show a dissociation between self-related emotional areas – such as the medial prefrontal cortex, that is activated when participants' rejections bear on their own payoff – and brain regions responsible for affective-motivational reaction to fairness norms' violations, such as the anterior insula, that is activated when rejections bear both on participants' payoff and on the others' (Civai et al. (2010b).

Social behavior is likely to be driven by a selected strategy that depends on a combination of automatic mechanisms as well as the environmental content and context. It is known that participants' performance in experimental games such as the Dictator Game or the Ultimatum Game (UG) is influenced by a wide variety of factors. For example, the degree of generosity in the dictators decreases together with the degree of anonymity towards both the receiver and the experimenter (Hoffmann et al. 1996); furthermore, Dana et al. (2007) have found that relaxing the transparency, that is, giving the dictator the illusion of fairness simply by increasing the uncertainty of the receiver's payoff, significantly decreases fair behavior. Other factors that influence preferences are the degree of self-involvement and intentions (Blount 1995; Falk & Fischbacher 2006; Güroğlu et al. 2010). In particular, an increase in the tolerance for unfair advantageous offers in the UG is predicted when the offers target self-payoff (Fehr & Schmidt 1999). As far as intentions are concerned, it has been widely demonstrated that the rejection rate for unfair UG offers decreases together with the perceived proposer's responsibility. An interesting norm-based model has been described by Bicchieri (2006), which stresses effects of framing on "people's expectations and perception of what norm is being followed" (Bicchieri & Zhang 2010, p. 18) affecting the final decision.

In light of these findings, social preferences cannot be considered as stable. They appear to be conditioned by the social situation, and they could be better defined as strategies, implemented in order to maximize (not in strict economical terms) the outcome. An experimental approach that investigates

the conditions that trigger the different strategies might explain the great variability that characterizes social behaviors such as reciprocity (e.g., Gneezy & Rustichini 2000). Theories of fast and frugal heuristics (Gigerenzer et al. 1999) might be successfully applied to the social mind (Hertwig & Herzog 2009). This would allow us to describe behavior as driven by fast and frugal social heuristics, such as "imitate the majority" or "group recognition," which, in turn, are triggered by different environmental factors.

To conclude, the issue of cooperation should be reconsidered in light of the fact that the different types of evidence discussed by Guala were collected under different experimental and nonexperimental conditions; hence, they are likely to reflect not a single process but different processes which, necessarily, must not be mutually exclusive.

Examining punishment at different explanatory levels

doi:10.1017/S0140525X1100121X

Miguel dos Santos and Claus Wedekind

Department of Ecology and Evolution, Biophore, University of Lausanne, 1015 Lausanne, Switzerland.

miguel.dossantos@unil.ch

http://www.unil.ch/dee/page57244.html

claus.wedekind@unil.ch

http://www.unil.ch/dee/page21538.html

Abstract: Experimental studies on punishment have sometimes been over-interpreted not only for the reasons Guala lists, but also because of a frequent conflation of proximate and ultimate explanatory levels that Guala's review perpetuates. Moreover, for future analyses we may need a clearer classification of different kinds of punishment.

When explaining behavioral decisions, it is important to distinguish between different explanatory levels, especially between proximate (mechanistic) and ultimate (evolutionary) explanations (Tinbergen 1963). Proximate explanations of a given behavior deal with questions about its ontogeny (e.g., how does the behavior change with age and experience) or about its causation, that is, the physiological, molecular, and cognitive mechanisms underlying the behavior and the stimuli that elicit it. Ultimate explanations either deal with questions about the phylogeny of the behavior (e.g., how does it compare with similar behaviors in related species) or its adaptive value (e.g., what is its impact on the individual's survival and lifetime reproductive success).

The concept of weak reciprocity, as defined in the target article, is an attempt to explain the adaptive value of cooperation and punishment because it concentrates on the fitness benefits one could get from cooperating, defecting, or punishing (Alexander 1974; Trivers 1971). This concept is restricted to one explanatory level only. In contrast, strong reciprocity mixes different explanatory levels: it uses proximate arguments to explain ultimate problems (Bowles & Gintis 2004; Fehr & Fischbacher 2003; 2004; Fehr & Gächter 2002; Fehr & Rockenbach 2003; Gintis et al. 2003). Strong reciprocity has been called, for example, a "predisposition to reward others for cooperative, norm-abiding behaviours" and "a propensity to impose sanctions on others for norm violations" (Fehr & Fischbacher 2003, p. 785). Such a definition clearly relates to the causal mechanisms of cooperation and punishment. But the concept is then frequently used as to answer ultimate (evolutionary) questions, for example, in Bowles and Gintis (2004, p. 17): "cooperation is maintained because many humans have a predisposition to punish those who violate group-beneficial norms." Such a mixing up of different explanatory levels can, from an evolutionary point of view, easily lead to over-interpretations of proximate patterns (Hagen & Hammerstein 2006; Rankin et al. 2009; Sigmund 2007; West et al. 2007a; 2011). For example, punishment that can be observed in anonymous one-shot interactions seems truly altruistic and was interpreted as such in Fehr and Gächter (2002). However, until very recently, humans lived in groups where anonymous one-shot interactions were probably very rare, that is, such interactions are most probably not the context in which human punishment has evolved. If studied within a more natural social context, human punishment may ultimately be self-interested.

As discussed in the target article, explaining punishment from an evolutionary point of view requires determining the costs and benefits of punishment. In line with weak reciprocity models, recent studies have shown that punishment can lead to long-term net benefits and hence be evolutionarily stable when punitive actions contribute to a punishment reputation (dos Santos et al. 2011; Hilbe & Sigmund 2010). Under such conditions, the immediate costs of punishment can be outweighed by the benefits a punisher receives later because of his or her punishment reputation. Experimental studies that ignore the possible effects of a punishment reputation can therefore easily produce artifacts (Hagen & Hammerstein 2006).

We also believe that the term "punishment" is currently used too broadly in the literature on cooperation. If punishment is the subtraction of resources from free- riders in order to reduce the frequency of further free-riding, there are at least three different kinds of punishment that may need to be distinguished both for ultimate and proximate analyses. Many of these analyses deal with what could be called "simple costly punishment"; that is, punishers pay a cost to induce a cost on the punished (Dreber et al. 2008; Fehr & Gächter 2000a; Rand et al. 2009a; Rockenbach & Milinski 2006; Wu et al. 2009).

Another form of punishment could be called "punishment by taking something away" (e.g., as in Cephu's story, described in the target article). Here, the punisher takes something from the punished in order to induce a cost to the punished. Regardless of whether or not the punisher thereby experiences an immediate reduction of his or her own welfare, "punishment by taking something away" and the above-mentioned "simple costly punishment" are likely to differ in their cost-benefit ratios (relevant for ultimate analyses) and may involve, for example, different kinds of emotions (relevant for proximate analyses).

A third category could be called "punishment by refusal." The punisher then punishes by refusing to cooperate with the punished in a repeated game like, for example, an iterated Prisoner's Dilemma (Fudenberg et al. 1994). The examples of ostracism discussed by Guala relate to this kind of punishment. Such defection may typically be a reaction to non-provoked defection and could be called "punishment" if it reduces the income of the punished (i.e., his or her benefits from what would otherwise be cooperative interactions) in order to possibly improve the punisher's long-term benefits from future cooperative interactions with a refined punished or with others.

This third kind of punishment could be immediately costly for the punisher, for example, if it delays the resumption of beneficial mutual cooperation. Such immediate costs would have to be compensated in the long run in order to maintain "punishment by refusal" as an evolutionary successful behavioral strategy. However, a possible alternative function of defection in response to defection may be to simply avoid the losses of anticipated further defection (e.g., avoiding the sucker's payoff in the Prisoner's Dilemma). It is probably not useful to call this latter form of defection "punishment" if it usually does not ultimately increase the level of cooperation within a group or directly with the defector (from an ultimate point of view), or if it is just a precautionary measure to avoid further losses (from a proximate point of view). Therefore, purely punitive actions may not always be easy to identify. Multidisciplinary approaches that

carefully exploit the specific advantages of proximate and ultimate analyses are therefore often necessary to better understand human behavior.

Retaliation and antisocial punishment are overlooked in many theoretical models as well as behavioral experiments

doi:10.1017/S0140525X11001221

Anna Dreber[a] and David G. Rand[b]

[a]*Department of Economics, Stockholm School of Economics, 113 83 Stockholm, Sweden;* [b]*Program for Evolutionary Dynamics, Harvard University and Department of Psychology, Harvard University, Cambridge, MA 02138.*

anna.dreber@hhs.se drand@fas.harvard.edu

http://sites.google.com/site/annadreber/

http://www.people.fas.harvard.edu/~drand/

Abstract: Guala argues that there is a mismatch between most laboratory experiments on costly punishment and behavior in the field. In the lab, experimental designs typically suppress retaliation. The same is true for most theoretical models of the co-evolution of costly punishment and cooperation, which a priori exclude the possibility of defectors punishing cooperators.

The target article is interesting and raises many important questions about the role of costly punishment in the evolution of human sociality. Particularly relevant is the mismatch between the design of many economic experiments and the conditions under which human evolution (genetic or cultural) seems likely to occur.

Nearly all economic game experiments exploring costly punishment are explicitly designed to suppress the opportunity for retaliation and feuds. These experiments reshuffle either groups or identities from round to round and report the amount of punishment received only in aggregate. Such design features are highly unrealistic, and serve to cast costly punishment in the most positive possible light. As Guala points out, costly punishment can be disastrous in more realistic experimental settings with truly repeated interactions where retaliation is possible (Denant-Boemont et al. 2007; Dreber et al. 2008; Nikiforakis 2008; Wu et al. 2009). Not only can cooperators punish defectors, but defectors can also punish cooperators (Cinyabuguma et al. 2006; Gächter & Herrmann 2009; 2011; Herrmann et al. 2008). Furthermore, positive reciprocity (rather than costly punishment) can effectively maintain cooperation when repetition or reputation are allowed (Dal Bó 2005; Dal Bó & Fréchette 2011; Fudenberg et al., in press; Milinski et al. 2001; 2002; Rand et al. 2009a; Rockenbach & Milinski 2006; Wedekind & Milinski 2000), although see Vyrastekova and van Soest (2008).

This same critique also applies to almost all evolutionary game theoretic models of costly punishment and cooperation. Many models have been proposed to demonstrate how costly punishment could promote the evolution of cooperation (Bowles & Gintis 2004; Boyd et al. 2003; Gintis 2000; Hauert et al. 2007; Nakamaru & Iwasa 2005; 2006; Sigmund et al. 2010, Traulsen et al. 2009; Wang et al. 2011). Yet virtually all of these models assume that only cooperators punish defectors. Punishment targeted at cooperators ("antisocial punishment") is excluded a priori.

We feel that evolutionary models do best to include the full range of combinatorially possible strategies (of a specified level of complexity) and then to ask which are favored by natural selection. If, instead, the strategy set is restricted to only include strategies that seem logical or desirable, this can greatly affect the outcomes and potentially be quite misleading. We note that this is not unique to models based on "strong reciprocity" but also applies to most models that do not invoke group selection.

Recent theoretical work has examined the effect of retaliation and antisocial punishment, and the results have not been promising for "altruistic" punishment. When the opportunity to retaliate is added to a model based on intergroup conflict, punishment is much less effective at promoting cooperation (Janssen & Bushman 2008). In such models, groups of cooperators outcompete groups of defectors; but within a single group, defectors outcompete cooperators. Costly punishment allows cooperative groups to keep out defectors (Boyd et al. 2003). But now the second-order free-rider problem arises: Cooperators who punish are at a disadvantage relative to cooperators who do not punish. When defectors are rare, this disadvantage is small and does little to undermine cooperation. But when retaliation is possible, this exacerbates the second-order free-rider problem: Not only do cooperative punishers bare the cost of punishing relative to non-punishing cooperators, but they also incur the additional cost of being retaliated upon. Thus, punishment has only limited power to promote cooperation.

Similar results are obtained when antisocial punishment (as well as indiscriminant punishment) are allowed in a model based on spatial structure (Rand et al. 2010). In this second model, interaction and competition only occur with those nearby, thus relative payoff is key and spite is adaptive. When only cooperators can punish defectors, punishment allows cooperation to dominate: Without punishment, cooperators are at a disadvantage because they pay the cost of cooperation; but by punishing defectors, they can regain the relative advantage (Nakamaru & Iwasa 2005; 2006). When all punishment strategies are available, however, defectors can also punish cooperators (a form of anticipatory retaliation). Thus, the prosocial and antisocial punishments cancel each other out, and punishment no longer allows cooperation to proliferate. Instead, the only strategy that is globally stable is to defect and then punish cooperators.

Likewise, punishment no longer promotes cooperation in an optional Public Goods game model once antisocial punishment is allowed (Rand & Nowak 2011). In optional cooperation games, defectors invade cooperators, loners that opt out of the game in favor of a fixed intermediate payoff invade defectors, and cooperators invade loners (Hauert et al. 2002). Allowing cooperators to punish defectors breaks this rock-paper-scissors cycle and stabilizes cooperation (Hauert et al. 2007). But when all punishment strategies are possible, the cycle is as easily broken by defectors that punish loners, or loners that punish cooperators. Antisocial punishment is common, and punishment does not substantially increase cooperation compared to a game without punishment.

Another model considers prosocial and antisocial punishment as well as retaliation in the context of a repeated Prisoner's Dilemma game (Rand et al. 2009b). A Nash equilibrium analysis finds many cooperative equilibria that pay to punish defection. Yet stochastic evolutionary simulations find that selection consistently disfavors costly punishment, and these simulations show quantitative agreement with a set of behavioral experiments (Dreber et al. 2008). In equilibrium, costly punishment is not actually costly – just the threat is sufficient to maintain cooperation. But in the noisy world of stochastic game dynamics, mutation and (relatively) weak selection lead to heterogeneous populations: The punisher must pay. Hence, costly punishment is disfavored, and instead evolution leads to traditional tit-for-tat strategies that "punish" defection not with costly punishment, but rather with denial of future reward.

Thus, the issues raised by Guala with respect to artificial experimental designs also apply to many evolutionary game theoretic models. Initial explorations of allowing retaliation and antisocial punishment in these models find the power of punishment for promoting cooperation to be much reduced or non-existent; further work in this vein is an important direction for future study.

Gossip as an effective and low-cost form of punishment

doi:10.1017/S0140525X11001233

Matthew Feinberg[a], Joey T. Cheng[b], and Robb Willer[c]

[a]*Department of Psychology, University of California, Berkeley, Berkeley, CA 94720-1650;* [b]*Department of Psychology, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada;* [c]*Department of Sociology, University of California, Berkeley, Berkeley, CA 94720-1980.*

matthewfeinberg@berkeley.edu

http://sites.google.com/site/matthewfeinbergpsychology/

joeycheng@psych.ubc.ca

willer@berkeley.edu http://willer.berkeley.edu/

Abstract: The spreading of reputational information about group members through gossip represents a widespread, efficient, and low-cost form of punishment. Research shows that negative arousal states motivate individuals to gossip about the transgressions of group members. By sharing information in this way groups are better able to promote cooperation and maintain social control and order.

Central to Guala's target article is the claim that experimental studies of costly punishment should not be interpreted as evidence for the existence of costly punishment outside the lab, but at best as evidence for the existence of strong *motivations* to punish those who have behaved antisocially. In most field settings, however, these motivations are likely to manifest in lower or zero-cost behaviors like ridicule, ostracism, and gossip. We agree with this point and highlight the specific role played by gossip as a ubiquitous form of low-cost punishment prevalent in all known human societies. Indeed, Dunbar (2004) estimates that gossip constitutes 65% of all spoken communication.

We argue that gossip – the sharing of evaluative information about an absent third party – is a widespread and highly effective form of punishment found in field settings (Dunbar 1996/1998) that alleviates the need for costlier forms of punishment. Gossip promotes cooperation in groups in two primary ways: (1) by spreading reputational information that warns group members about a transgressor, leading them to avoid or ostracize the transgressor; and (2) by increasing reputational incentives that deter individuals from behaving antisocially (Beersma & van Kleef, in press; Feinberg et al. 2011; Willer et al. 2010).

Recent research finds that the social psychological dynamics driving gossip correspond quite well with the motives revealed by experimental research on costly punishment. In a series of studies, Feinberg et al. (2011) demonstrate that gossip is driven by the same negative affective response that underpins costly punishment. After witnessing a target behave selfishly in a social dilemma situation, observers showed heightened levels of negative affect (e.g., frustration, annoyance) and physiological arousal, both of which were reduced by passing on reputational information to the transgressor's future interaction partner. A subsequent study showed that participants would gossip even when it required investing their own earnings to do so. Akin to altruistic punishment findings, these results suggest that when individuals detect the presence of defectors in the environment, they experience a strong motivation to share reputational information with other group members, even when doing so is costly. Additional research has found gossip deters antisocial behavior; when given the opportunity to behave selfishly in a social dilemma, individuals behaved more prosocially if they knew an observer was likely to gossip about them (Beersma & van Kleef, in press; see also Dunbar 1996/1998; 2004; Piazza & Bering 2008a; Sommerfeld et al. 2007).

Whereas Guala emphasizes that the anthropological evidence fails to show robust patterns of costly punishment in the field, there is substantial cross-cultural evidence for the prevalence of gossip outside the lab. Evidence that gossip serves as a mechanism for maintaining cooperation has been demonstrated in small societies in Mexico, Polynesia, and Fiji, to name a few (Arno 1980; Besnier 1989; Haviland 1977). It is sensible that gossip would be so widely used in small egalitarian societies because of its efficiency, effectively promoting cooperation at minimal cost. The small size of these societies means that all members know one another, ensuring that information can potentially spread to all group members and recipients of gossip know and potentially interact with the target. Additionally, in small societies, the spread of negative reputational information has a significantly greater impact on transgressors, with each individual person hearing of one's negative reputation representing a larger proportion of the group aware of the transgression. Moreover, gossip's low cost alleviates potential second-order free-rider problems that more costly punishment behaviors typically face.

Because of its effectiveness and low cost, we should expect gossip to be a more common response to the observation of antisocial behavior than more costly forms of punishment. This notion is consistent with evidence suggesting that costly punishment may become limited in environments where indirect reciprocity or reputational information offers a cheaper means of social control (Rockenbach & Milinski 2006). That said, the fact that gossip is a more efficient tool of punishment in most settings does not rule out the possibility of more costly punishment in situations where gossip is impractical or ineffective.

Guala views gossip as a costless form of punishment, and we agree that its low-cost nature is likely critical to its prevalence. But the costs and benefits of gossip remain unclear and deserve future study. Gossip entails risks of retaliation and reputation loss. At the same time, it is also possible that gossip could offer benefits to the gossiper (Willer 2009; Willer et al. 2010). Passing on reputational information may lead to a variety of possible benefits: (1) deterring antisocial behavior directed towards the gossiper by communicating that he or she will readily spread information about antisocial behavior; (2) improving status by advertising the extensiveness of the gossiper's connections in the group's social network (Cheng et al. 2007); and (3) advertising the gossiper's prosociality, thereby making him or her an attractive, trustworthy partner. Future research is needed to better understand the magnitude of costs and benefits associated with gossip and how these might vary across different contexts.

Blood, sex, personality, power, and altruism: Factors influencing the validity of strong reciprocity

doi:10.1017/S0140525X11001245

Eamonn Ferguson[a] and Philip Corr[b]

[a]*Personality and Social Psychology Group, School of Psychology, University of Nottingham, Nottingham, NG7 2RD, United Kingdom:* [b]*School of Social Work and Psychology, and Centre for Behavioural and Experimental Social Science (CBESS); University of East Anglia., Norwich NR4 7TJ, United Kingdom*

eamonn.ferguson@nottingham.ac.uk

http://www.psychology.nottingham.ac.uk/staff/ef/home.html

P.Corr@uea.ac.uk

http://www.ueapsychology.net/differential-psychology-pg14.html

Abstract: It is argued that the generality of strong reciprocity theory (SRT) is limited by the existence of anonymous spontaneous cooperation, maintained in the absence of punishment, despite free-riding. We highlight how individual differences, status, sex, and the legitimacy of non-cooperation need to be examined to increase the internal and ecological validity of SRT experiments and, ultimately, SRT's external validity.

In his critique of strong reciprocity theory (SRT), Guala highlights some concerns with its external validity, but contends that its internal validity is less problematic. We endorse the concerns about external validity, but raise additional concerns with respect to internal validity. We suggest ways to improve the

ecological validity of laboratory-based studies in order to enhance their external validity.

External validity – Cooperation without punishment. Guala argues that the key source of disagreement concerning the external validity of SRT is whether or not costly punishment, in the face of free-riding, supports spontaneous cooperation outside the laboratory. Guala argues that punishment to support cooperation is, in fact, coordinated and cheap. However, it should be acknowledged that there are many forms of spontaneous cooperation that emerge in the absence of punishment, despite free-riding. One example is voluntary blood donation. The donor and recipient remain anonymous and never meet. Blood donors tend not to talk about being a donor (Ferguson & Chandler 2005), and the number of donors they know does not influence their decision to donate (Piliavin & Callero 1991). While there are organized blood drives, blood donors tend to donate at drop-in centres at times they find convenient. Although an anonymous, relatively high-cost spontaneous act of altruism, blood donation is marked by a large free-rider problem: about 6% of the eligible population donate (Ferguson et al. 2007). Evidence shows that feelings of warm glow (Andreoni 1990) are a key motivation for blood donation (Ferguson et al. 2008; in press) Recent evidence also shows that registering for posthumous organ donation is likewise motivated by emotional regulation through anticipated regret (O'Carroll et al. 2011). Finally, under specific conditions free-riders may be tolerated as this enables maximum group benefit (MacLean et al. 2010). Thus, punishment is not needed for many forms of cooperation.

Internal validity – Individual differences. An experiment is internally valid when alternative explanations or additional mechanisms that contribute to the effect are identified or controlled. Given the degree of heterogeneity observed in economic tasks, there is a growing realization that economic models need to consider the role of personality traits as additional explanatory variables (Ferguson et al. 2011; Wischniewski et al. 2009). Ferguson et al. (2011) propose a model of personality within economics whereby the expressed behaviour on any economic task reflects the motivations associated with the personality trait, as well as task constraints, incentives, and so forth. For example, someone who is motivated to maximize rewards is more likely to show cooperative behaviour when reputation building is possible, but free-ride in an anonymous game: behavioural expression is strategic.

One implication for SRT is that some agents will punish only when it is strategically advantageous to do so. Indeed, this is what has been observed with respect to those high in Machiavellianism, who behave selfishly when punishment is not expected, but cooperate when it is (Spitzer et al. 2007). Individual differences will also influence decisions both to punish and how to respond to punishment. Some individuals (e.g., psychopaths) will be more willing than others to punish across all contexts; however, some others (e.g., the highly anxious) will respond to punishment by cooperating, while others will continue to free-ride or retaliate (e.g., psychopaths). Finally, when individuals have the opportunity to meet and communicate (as is the real world), some (e.g., those high in psychopathy) will be more likely to exploit people whom they identify as possessing exploitable traits (e.g., agreeableness; Buss & Duntley 2008). Thus, while expressed behavioural styles are often noted in the SRT literature (Ule et al. 2009), these may reflect, in part, the operation of pre-existing individual differences.

Ecological validity. By increasing the ecological validity of laboratory-based tests of SRT, the external validity gap can be reduced (List 2009). Guala highlights studies allowing for multiple punishment options, communication, or retaliation. We highlight two other evolutionary parameters that are important in this regard: (1) resource holding and status and sex; and (2) legitimacy of free-riding.

While resources (e.g., physical, financial) are not equally distributed in the population, this is not captured within standard laboratory tests of SRT, where participants often have access to equal resources. The unequal distributions of resources will influence levels of aggression (punishment). For example, people are less likely to be aggressive towards those with greater physical resources (Archer & Benson 2008). Indeed, people are less willing to punish a transgression by high-status individuals or groups (Eckel et al. 2010). There are also established sex differences in the use of aggression: Males use direct aggression, and women indirect (Archer 2004). Therefore, pronounced sex differences should be observed in both the degree and type of punishment used in laboratory tests of SRT. Women also may be less willing to punish males who hold stronger physical resources. Thus, power, status, and sex may result in the use of punishment for many different reasons other than to enforce norms of fairness.

In real-world contexts, non-contributions may occur for a number of legitimate reasons (e.g., illness), and for sustained cooperation people need to be able to distinguish legitimate from non-legitimate non-cooperation (Lotem et al. 1999). Within laboratory tests of SRT, free-riders do not have any legitimate reason not to contribute (all have an endowment and can contribute). SRT, therefore, needs to examine the role of legitimacy of non-contribution and how this influences the level of punishment adopted. Legitimate free-riders should be treated like cooperators.

To conclude, the generality of SRT is limited by the existence of anonymous spontaneous cooperation, in the face of free-riding, that is maintained in the absence of punishment or even conditions where free-riding is tolerated. Personality, status, sex, and legitimacy of non-cooperation need to be examined in order to increase the internal and ecological validity of SRT experiments and, ultimately, its external validity.

In the lab and the field: Punishment is rare in equilibrium

doi:10.1017/S0140525X11001415

Simon Gächter

School *of Economics, University of Nottingham, University Park, Nottingham NG7 2RD, United Kingdom.*

simon.gaechter@nottingham.ac.uk

http://www.nottingham.ac.uk/Economics/people/simon.gaechter

Abstract: I argue that field (experimental) studies on (costly) peer punishment in social dilemmas face the problem that in equilibrium punishment will be rare and therefore may be hard to observe in the field. I also argue that the behavioral logic uncovered by lab experiments is not fundamentally different from the behavioral logic of cooperation in the field.

Francesco Guala's target article is a valuable contribution to the discussion about the empirical importance of weak and strong reciprocity. In this commentary I focus on one aspect of his call for more field (experimental) evidence on strong negative reciprocity in social dilemmas. I argue that collecting such evidence is welcome but faces the difficulty that theory predicts, and experiments confirm, that punishment will be rare in equilibrium. I also argue that from the equilibrium perspective the behavioral logic uncovered by lab experiments is not fundamentally different from the behavioral logic of cooperation in field situations. Therefore, the distinction between "narrow" and "wide"

readings of strong reciprocity and the preoccupation with external validity concerns is somewhat artificial.

I agree with Guala that experiments are a good tool to measure motivations. Previous experimental evidence (surveyed in, e.g., Chaudhuri 2011; Gächter & Herrmann 2009) shows that many people are willing to incur costs to punish freeloaders. People punish in finitely repeated games played by the same set of people (e.g., Fehr & Gächter 2000a), in repeated one-shot experiments where people interact with new group members each time (e.g., Fehr & Gächter 2002), and even in single-shot experiments where group members interact exactly once (e.g., Cubitt et al. 2011. The main purpose of these experiments was to probe whether and under what conditions people are willing to incur costs to punish and whether this influences cooperation levels. For this purpose and for various historical (and logistical) reasons, almost all experiments implemented at most ten rounds of interaction. Although short experiments can show that punishment exists and can have powerful behavioral consequences, these kinds of experiments might not be long enough to allow equilibration to be observed. In equilibrium people will have shared expectations about how others will behave and will adopt their behavior accordingly; punishment should be rare. Short experiments make an equilibrium perspective difficult and may therefore "overstate" the frequency of punishment.

An equilibrium perspective is important, however, if one is interested in field (experimental) evidence on punishment. In many field settings, more or less stable groups of people will interact, and/or people will have more or less settled expectations (through own observation and experience, as well as through social learning) about how others will behave even in one-shot settings. In terms of observing punishment for freeloading in social dilemmas, theory predicts that in equilibrium punishment will be rare, even if some people are prepared to punish freeloaders. In the presence of punishers, freeloaders have an incentive to contribute to the public good to avoid punishment. Thus, if punishment is behaviorally effective and there is no antisocial punishment of cooperators (Herrmann et al. 2008), punishment will be rarely used because there will be only few selfish transgressions (see also Boyd et al. 2003). As a consequence, the costs of punishment can be low.

The experimental evidence reported in Gächter et al. (2008) supports this reasoning. There are two conditions in their experiment: one without punishment, and one with punishment. In both conditions, groups of three people each receive an endowment of 20 tokens which they can contribute to a public good or keep for themselves. Payoffs are such that people have an incentive to contribute nothing to the public good, although full contribution is the socially optimal decision. In the experiment without punishment, a round ends after everyone has made his or her contribution decision. In the experiment with punishment, a second stage is added where group members are informed of each others' contribution and then can decide to incur their own costs to reduce each others' earnings from the first stage by three money units. Punished group members are only informed of the sum of received punishment, and not about who punished them (in this sense punishment is "coordinated"). Group members interact for 50 periods, which should give plenty of time to allow for equilibration. Seventeen groups each participated in the two conditions (between subjects). Figure 1 reports the most important result for the purpose of this comment.

In the absence of punishment (labeled "Contributions N"), contributions start at 9.5 tokens on average and decline to 3.7 tokens by period 50. The average contribution in the second half of the experiment is 6.4 tokens. Adding the punishment opportunity has huge consequences on average contributions (labeled "Contributions P"). Contributions increase rapidly to 17.6 tokens on average in the second half and are significantly higher than in the first half ($n = 17$, $z = 3.39$, $p = 0.0007$; Wilcoxon signed ranks test with group average contributions as independent observations).

Most importantly for my present purposes, the dotted line illustrates the average frequency of punishment acts across the 50 periods (measured on the right-hand axis). Because each group consists of three members, each subject in each period has two opportunities to punish other group members. Thus, in each period each group has six punishment opportunities. Because there were 17 groups, the total number of possible punishment acts was $17 \times 6 \times 50 = 5{,}100$. Across the whole experiment we observe 493 acts of punishment (i.e., in 9.7% of all possible cases). Punishment was significantly more frequent in the first half than in the second half of the experiment (14.2 vs. 5.1%; $n = 17$, $z = 3.29$, $p = 0.001$, Wilcoxon signed ranks test).

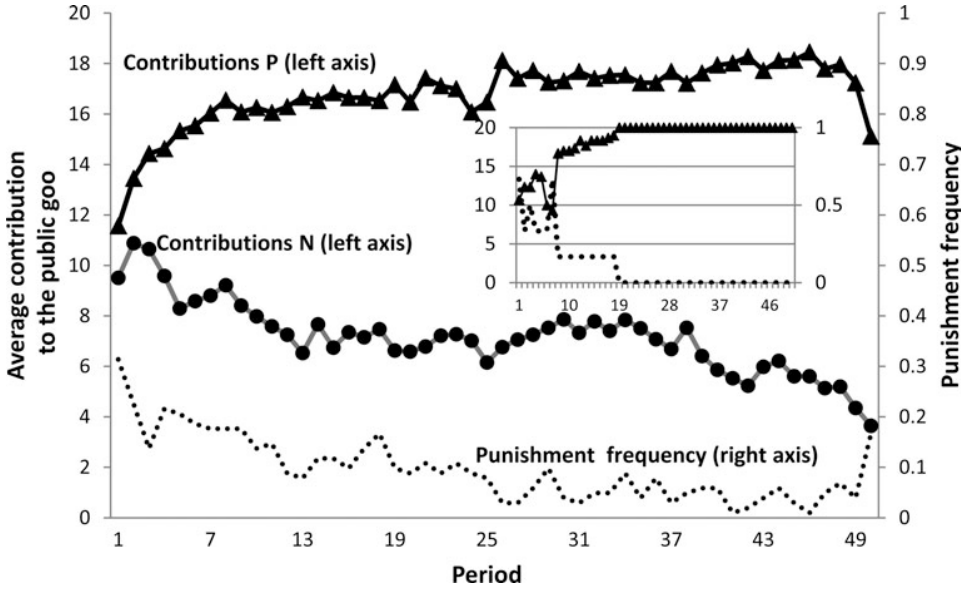

Figure 1 (Gächter). Average contributions to the public good of 17 three-person groups each across 50 rounds in the punishment condition (P) and the no punishment condition (N). Both are measured on the left axis. The dashed line depicts the frequency of punishment acts (measured on the right axis). The inlet figure illustrates contributions and punishment frequency of the median group with regard to cooperation level. Data are taken from Gächter et al. (2008); analysis and illustration are my own.

The inlet figure illustrates the median group with regard to contributions to the public good in the P experiment. This median group contributed 18.4 tokens on average and punished in 10% of all cases. Punishment occurred exclusively in the beginning of the experiment. From period 19 onwards not a single act of punishment was observed; and 100% of all contributions were maximal. Across all 17 groups punishment frequency and contribution level are significantly negatively correlated (Spearman rank, $n = 17$, $\rho = -0.75$, $p = 0.0005$). In the second half this correlation is $\rho = -0.93$, $p = 0.0000$).

My preferred interpretation of these results in the present context is that the first part of the experiment is a phase where common expectations are established, and behavior has to be coordinated accordingly. Once behavior and expectations are coordinated, equilibration has occurred and punishment is only rarely needed. The initial phase of the experiment might be seen as "artificial," but since the experiment is a novel environment for participants, this is an inevitable part and the lab analogue of social learning in the field. In the field, social learning and cultural transmission (learning from peers, teachers, parents) teach people what constitutes socially acceptable behavior and what gets punished (e.g., Henrich 2004). Once expectations are formed and behavior is adapted accordingly, the same behavioral logic holds in the lab and the field, despite obvious environmental and complexity differences between them. To criticize strong reciprocity theorists for their "narrow focus on artificial environments" (target article, sect. 15, para. 2) therefore seems to miss the point.

Think of queuing as an example. Orderly queues are socially optimal, but individual incentives are to jump the queue. Many queues are surprisingly orderly and few instances of punishment occur (counterexamples exist, of course). People learn through observation and education how to behave in a queue. Chances are that a queue jumper will be told off and sent back to line (which is an example of peer punishment). Many potential queue jumpers will think about this possibility and refrain from jumping the queue, making punishment a rare event. Thus, the behavioral logic of how to behave in a queue is the same as in the lab experiment. Sometimes queue jumpers go unpunished and queues break down, like in the lab where cooperation sometimes also works badly, despite, or because of, punishment (Herrmann et al. 2008).

The relevance of this example extends beyond queuing. Peer punishment as modeled in these experiments can be seen as expressions of social disapproval (Carpenter & Seki 2011; Masclet et al. 2003), which are ubiquitous in social life (think of ridiculing, gossiping, reprimands, social exclusion, etc.). Disapproval will often be costly for the sanctioned individual and in many cases will also be costly to the punisher, in terms of psychic costs (at least some people find it difficult to confront wrongdoers), foregone opportunities, and possible retribution. Therefore, one should not interpret these experiments too narrowly in terms of direct material costs alone. Modeling punishment in material terms is primarily done to control for individual incentives and to allow for exact theoretical predictions (Smith 1982). The behavioral logic of meting out and avoiding punishment in the lab is similar to disapproving of some people's behavior and avoiding disapproval. Peer punishment experiments should therefore be seen at least as much as models of social control or moralistic aggression than of direct material punishment.

The experiment also suggests that the lack of observing punishment in field contexts cannot be taken as evidence for the irrelevance of punishment and as a sort of lab artifact. The experiment shows that even occasional punishment can have a huge impact on pro-social behavior as compared to a situation where people know for sure that they can get away with selfish behavior. Further experiments that model other important aspects of reality, like the possibility of communication (e.g., Bochet et al. 2006), coordinated punishment (e.g., Casari & Luini 2009), third-party punishment (Fehr & Fischbacher 2004), assortative matching (e.g., Gächter & Thöni 2005), or the simultaneous presence of rewarding strategies (e.g., Rockenbach & Milinski 2006; Ule et al. 2009), also suggest that punishment will be rare in equilibrium and nevertheless have an important behavioral impact.

Finally, the equilibrium perspective suggests it is at least as important to focus on evidence about the punishment people *expect* would they transgress, rather than actual punishment, as well as on institutions like monitoring that might result in punishment (Rustagi et al. 2010).

The social structure of cooperation and punishment

doi:10.1017/S0140525X11000914

Herbert Gintis[a] and Ernst Fehr[b]

[a]*Santa Fe Institute, Santa Fe, NM 87501;* [b]*Department of Economics and Laboratory for Social and Neural Systems Research, University of Zurich, CH-8006 Zurich, Switzerland.*

hgintis@comcast.net efehr@iew.unizh.ch

Abstract: The standard theories of cooperation in humans, which depend on repeated interaction and reputation effects among self-regarding agents, are inadequate. Strong reciprocity, a predisposition to participate in costly cooperation and the punishment, fosters cooperation where self-regarding behaviors fail. The effectiveness of socially coordinated punishment depends on individual motivations to participate, which are based on strong reciprocity motives. The relative infrequency of high-cost punishment is a result of the ubiquity of strong reciprocity, not its absence.

Standard models of human cooperation in economics and biology assume purely self-regarding agents who use repeated interactions (reciprocal altruism) or public reputations (indirect reciprocity) to sustain mutual helping behaviors. While these mechanisms are important, there are many equally important forms of prosocial behavior which cannot be accounted for in the same way (Bowles & Gintis 2011; Fehr & Gintis 2007; Gintis 2005; 2009). These include: voting in elections, participating in collective actions, being kind to strangers, contributing to community public goods, and behaving morally in anonymous situations, or where the material penalties for immoral behavior are low.

Economic experiments strongly suggest that human prosociality is not limited to calculated selfishness (e.g. Batson 1991; Fehr & Gächter 2000a; 2000b; 2002; Fehr et al. 1997), but that the presence of free-riders is a key and ever-present threat to sustained cooperation. *Strong reciprocity*, a behavioral mechanism including both altruistic cooperation and costly punishment of free riders (Gintis 2000) thus helps sustain cooperation over long periods. This work showed that humans have strong and consistent other-regarding preferences that could be enlisted in support of social cooperation. In fact, anthropologists have confirmed that strong reciprocity is indeed routinely harnessed in the support of cooperation in small-scale societies (Boehm 1984; 1999; Henrich et al. 2010a; Wiessner 2005; 2009), as stressed in Henrich & Chudek's commentary in this issue.

Guala characterizes the punishment side of strong reciprocity as "uncoordinated." This is simply incorrect. Collective action is a real-life expression of strong reciprocity (Bowles & Gintis 2004, p. 17), and the predisposition to punish "transgressors" is often socially organized and sanctioned. Indeed, individuals are often deterred from carrying out self-initiated sanctions (Boyd et al. 2010). The experimental evidence for coordinated punishment was laid out in several experimental papers on strong reciprocity (e.g., Cinyabuguma et al. 2005).

Guala claims that costly punishment is rarely observed in the real world, and what punishment is observed is generally not very harsh (e.g., verbal harassment, gossip, ostracism). These observations, even if true, in no way conflict with strong reciprocity models of social cooperation. First, if punishment is effective, it will be rarely carried out. Thus, the absence of frequent punishment is an indication that the threat of punishment has a particularly strong effect. For instance, the average taxpayer in the United States is never penalized for tax evasion, yet no one doubts the importance of prosecuting tax evasion. Similarly, most drivers receive only a few traffic citations in the course of their lives, but many drivers adjust their driving to avoid citations. Second, we stress that most humans are very averse to public criticism of even a verbal form of punishment, and we cite studies that show that verbal criticism alone often leads to conformity (Masclet et al. 2003). In addition, the human emotion of shame serves to amplify social criticism, thereby lessening the need for costly punishment (Bowles & Gintis 2005; Gintis 2004). Moreover, Guala seriously understates the importance of diffuse, uncoordinated, costly punishment in promoting norm adherence.

Guala claims that some punishment is "zero cost." If so, this would add an interesting dimension to the strong reciprocity model, but it does not conflict with this model.

In sum, we agree with Guala that socially structured punishment is important, but we assert that the predisposition to reward goodness and punish evil underlies the effectiveness of socially structured punishment. We also reaffirm the critical importance of diffuse, unstructured cooperation and punishment in fostering social efficiency and a high quality of life.

Is strong reciprocity really strong in the lab, let alone in the real world?

doi:10.1017/S0140525X11001257

Şule Güney and Ben R. Newell

School of Psychology, University of New South Wales, Sydney, NSW, 2052, Australia.

s.guney@unsw.edu.au

http://www.psy.unsw.edu.au/profiles/phd/sguney.html

ben.newell@unsw.edu.au

http://www2.psy.unsw.edu.au/Users/BNewell/

Abstract: We argue that standard experiments supporting the existence of "strong reciprocity" do not represent many cooperative situations outside the laboratory. More representative experiments that incorporate "earned" rather than "windfall" wealth also do not provide evidence for the impact of strong reciprocity on cooperation in contemporary real-life situations or in evolutionary history, supporting the main conclusions of the target article.

The core phenomenon discussed in the target article is strong reciprocity: a predisposition to reward cooperators (altruistic rewarding) and punish norm violators (costly punishment), at a personal cost, even when the probability that this cost will be repaid is very low (Gintis 2000; Gintis et al. 2003).

Guala calls our attention to the fact that the existence of strong reciprocity appears to be strongly supported by *laboratory experiments*; yet there is no evidence from *outside the laboratory* supporting the claim that strong reciprocity (especially costly punishment) sustains cooperation in real life.

We agree with Guala's point that evidence gathered from controlled laboratory experiments is not sufficient to reach the conclusion that strong reciprocity sustains cooperation in real life. However, we further argue that it is not only the absence of real-life data that is problematic for strong reciprocity theorists. There also exist *laboratory experiments* which present strong *counter-evidence* against the presence of strong reciprocity and the claims for its evolutionary origins.

The significance of observing a certain behavior pattern (strong reciprocity) in the laboratory for understanding a real-life phenomenon (cooperation) depends crucially on the extent to which the real-life phenomenon is appropriately captured in an experimental setup. Consider a real-life situation in which a group of individuals engages in a cooperative act, say, hunting a stag. In an experimental setting this corresponds to an economic game (e.g., Ultimatum Game, Public Goods game, or Trust game) played out by a group of experimental participants. In real life there needs to be something that motivates individuals to engage in this cooperative act. For the hunters, this could be the idea of satisfying their hunger. In the laboratory one can try to elicit this effect with the use of financial incentives – so far, so good.

However, in order to satisfy their hunger, the hunters have to make a plan, then run after the stag, then use their tools or weapons to kill the stag, then cut it into pieces and distribute them amongst the group. In short, the hunters have to put in some *real effort* to get this material benefit. This is where real life and standard laboratory experiments diverge: In the laboratory, experimental subjects are generally *endowed* with an amount of money by the experimenter. They sit on a chair, make a decision, push a button on the keyboard, and leave the experiment with some money that they "earned" without putting in any actual effort. In the hunting scenario this would be like stumbling upon a recently deceased stag and simply having to decide how to share the spoils.

Importantly for the arguments developed in the target article, experiments which incorporate this naturalistic property of expending real effort to earn wealth or status show behavior that clearly diverges from the predictions of strong reciprocity. For instance, in one-shot, anonymously played Dictator games, a huge proportion of the allocators give tiny amounts or even zero to the receivers when the *allocators have earned* the amount to be distributed (Cherry et al. 2002). However, when the *receivers have to put in effort* to earn their share, the allocators give more than half of the pie to the receivers (Oxoby & Spraggon 2008). Similarly, in an Ultimatum Game, the proposers offer less to the responder when the proposers have to earn their status and wealth than when they are simply allocated to their roles and endowed with money by the experimenter (Hoffman et al. 1994).

These examples constitute a problem for the claim that strong reciprocity is an important mechanism for sustaining cooperation outside the laboratory because windfall wealth is much rarer in real life than earned wealth. If one's theory only applies to the minority of real-life situations, then to conclude that costly punishment, a component of strong reciprocity, sustains cooperation in real life appears unwarranted. Put simply, including consideration of such "earned-wealth" experiments strengthens Guala's conclusions about the impact (or lack thereof) of altruistic rewarding and costly punishment on the emergence of cooperation in evolutionary history.

Understanding the research program

doi:10.1017/S0140525X11001397

Joseph Henrich[a] and Maciej Chudek[b]

[a]*Department of Psychology and Department of Economics, University of British Columbia, Vancouver, B.C., V6T 1Z4, Canada;* [b]*Department of Psychology, University of British Columbia, Vancouver, B.C., V6T 1Z4, Canada.*

joseph.henrich@gmail.com **maciek@interchange.ubc.ca**

http://www2.psych.ubc.ca/~henrich/home.html

Abstract: The target article misunderstands the research program it criticizes. The work of Boyd, Richerson, Fehr, Gintis, Bowles and their collaborators has long included the theoretical and empirical study of models both with and *without* diffuse costly punishment. In

triaging the situation, we aim to (1) clarify the theoretical landscape, (2) highlight key points of agreement, and (3) suggest a more productive line of debate.

The target article muddles current theoretical issues regarding the evolution of human cooperation, and in the process creates an empty set of "strong reciprocity theorists." To begin, it makes little sense to oppose weak versus strong reciprocity. Weak reciprocity is a particular class of *theoretical* evolutionary models. By contrast, "strong reciprocity" is a label and summary description for a set of *empirical* regularities that emerged from work in the United States and Europe (it is not an evolutionary theory). To explain these regularities as well as much ethnographic and cross-cultural evidence (Henrich et al. 2004) – which are not well handled by weak reciprocity theories (Chudek & Henrich 2010, Fehr & Henrich 2003) – Boyd, Richerson, Fehr, Gintis, and Bowles (Guala's "strong reciprocity theorists," hereafter BRFGB) and others have proposed a wide range of cultural and genetic evolutionary models. These models represent hypotheses about what the important mechanisms might be that sustain social norms. In particular, much work has focused on understanding the *various way*s that cultural evolution can *harness* and *extend* aspects of our evolved psychology (e.g., kin psychology) to create stable prosocial norms that could be favoured by cultural group selection (Alvard 2003, Henrich et al. 2010a, Richerson & Boyd 1998). Only a subset of these models involve the diffuse costly punishment (DCP) referred to by Guala and observed in some public goods games.

Numerous contributions from BRFGB illustrate that they are in no way wedded to DCP. A decade ago, Gintis et al. (2001) modelled how signalling (not DCP) could favour the provision of public goods in a manner aimed at explaining ethnographic observations among the turtle-hunting Meriam (Smith et al. 2003). In 2004, Panchanathan and Boyd showed how cultural evolution could stabilize norms by linking a dyadic helping game to a public goods game (Panchanathan & Boyd 2004). There is no punishment there, let alone DCP. In his *Nature* perspective on this, Fehr (2004) emphasizes the importance of reputational mechanisms – not based on DCP – to stabilize social norms. In 2010, Boyd et al. showed how signalling can coordinate punishment to stabilize social norms. Guala mentions this paper approvingly, while not noticing that it is written by the *same authors* that he says adhere *only* to DCP.

In these models, as in all models involving DCP, sanctioners have higher payoffs/fitness than non-sanctioners (under the appropriate conditions). There is no magic; these are evolutionary models that explore which strategies are favoured by selective processes and under what conditions. Guala seems to suggest that models based on DCP require that punishment have a net long-term cost. That is false. Sometimes sanctioning costs are "paid-for" via inter-group competition (Boyd et al. 2003, Guzmanet al. 2007), and sometimes these systems are just mutually self-reinforcing (Boyd & Richerson 1992). Such costs have to net out somewhere, or be borne by some plausible constraint (weak reciprocity models all exploit mutational constraints; see Henrich (2004). Much evolutionary modelling has sought to identify how informal institutions (sets of norms or reputational systems) might reduce or eliminate these costs. Cultural group selection will often favour those mechanisms that more effectively incentivize the sanctioning of prosocial norms while sustaining internal harmony (Henrich et al. 2010a).

On the empirical side, the best evidence against the species-wide importance of DCP comes from a large-scale comparative project, initiated under the leadership of BFGB (Richerson not participating here), involving both experiments and ethnography in 24 different small-scale societies. Phase II of this project showed that community size is strongly positively associated with costly punishment. The analysis reveals that communities below a size of about 50 engage in little or no costly punishment (Henrich et al. 2010a, Marlowe et al. 2008). It is also the case that people from larger ethnic groups punish more. Going back a decade, these results confirm findings from Phase I of the project, in which people from three small-scale societies refused to reject low offers in the Ultimatum Game (Henrich 2000, Henrich et al. 2001).

The approach in this work not only explains the absence of DCP in many of the smallest-scale societies, it also accounts for why such punishing motivations emerge in larger-scale societies. Measures of punishment have not only been strongly associated with the size of stable communities and the success of ethnic groups, but they correlate strongly with gross domestic product (GDP) per capita across nations and with norms of civic cooperation and the rule of law (Herrmann et al. 2008). In Ethiopia, measures of conditional cooperation, which have been closely linked to individuals' willingness to punish, predict monitoring and effective commons management (Rustagi et al. 2010). Once properly theorized, behavioural game measures – including those related to punishment – readily link to real-world sociality (Henrich et al. 2010b).

While we agree that models relying on DCP (e.g., Henrich & Boyd 2001) are not consistent with how norms are actually stabilized in small-scale societies, a variety of other kinds of informal sanctioning mechanisms are, including those that coordinate or incentivize ostracism and punishing. Thus, when Guala pulls quotations from BRFGB's work that refer to "punishment," he implies that BRFGB refer only to DCP. If instead BRFGB mean "punishment" broadly defined, which results from community condemnations of norm violations channelled through local informal institutions (e.g., kinship or reputational systems), then Guala has misunderstood. Not only have BRFGB provided and promoted theoretical models not involving DCP, they also helped lead the project that have assembled the best evidence against the importance of DCP in small-scale societies. If by "punishment" they meant only DCP, then they would have to have been implicitly dismissing (1) some of their own theoretical models and (2) the fruits of their own anthropological collaboration. This seems unlikely.

We also agree with Guala's "narrow interpretation" of experiments, which was a central methodological element in the aforementioned large-scale comparative project. In the Phase I synthesis, one of the major points was that people bring motivations (values or heuristics) into the experimental games from everyday life (Henrich et al. 2005). The authors drew on interviews, ethnographic observations, and local history to interpret their experimental results.

Much of the research program pursued by BRFGB converges with that favoured in the target article, as a central thrust is to understand the origins of the formal and informal institutions that govern life in both small-scale and modern societies. However, progress might be better pursued by focusing on points of actual difference, such as why relying on fully specified evolutionary models is preferable to invoking the folk theorem (where many equilibria are dynamically unstable), or why it is crucial to consider how formal institutions interface with social norms to endow people with internalized motivations (Chudek & Henrich 2010).

Social preference experiments in animals: Strengthening the case for human preferences

doi:10.1017/S0140525X11001269

Keith Jensen

Research Centre for Psychology, School of Biological and Chemical Sciences, Queen Mary, University of London, London E1 4NS, United Kingdom.

k.jensen@qmul.ac.uk

http://www.sbcs.qmul.ac.uk/staff/keithjensen.html

Abstract: Guala appears to take social preferences for granted in his discussion of reciprocity experiments. While he does not overtly claim that social preferences are only by-products that arise in testing environments, he does assert that whatever they are – and how they

evolved – they have little value in the real world. Experiments on animals suggest that social preferences may be unique to humans, supporting the idea that they might play a prominent role in our world.

Guala's primary contention is that experimental evidence of reciprocity, particularly negative reciprocity, does not necessarily reflect what happens in the real world. This criticism of the "wide reading" of strong reciprocity does not challenge the "narrow reading," which suggests that social preferences motivate cooperation and punishment in the lab. In making his case, Guala appears to take social preferences for granted, and he does not wonder where they came from and what purpose they serve. Looking at other species will tell us something about the evolutionary history of social preferences and their possible adaptive value.

It is unlikely that the appearance of social preferences in the lab is artifactual. Concern for the welfare of others is one of the mechanisms that motivate our social behaviours. Much has been written about the role of empathy in motivating altruistic acts (Batson 1991). Feeling sad at the misfortunes of others and sharing their joy will provide short-term emotional benefits for acts that are otherwise costly when they are performed. Tangible benefits such as reciprocity or reputation are not the primary goals of the altruist. The same case can be made for the other "fortunes-of-others" emotions (Ortony et al. 1988), namely, jealousy (unhappiness at the good fortunes of others) and schadenfreude (pleasure in the misfortunes of others). These emotions can bridge the motivational gap between punitive and spiteful acts and any delayed or indirect benefits (e.g., reputation). As Guala points out, the laboratory environment allows for the detection of these preferences. It is possible that in the field, other preferences (such as for reputation) and alternative options (such as ostracism and other less costly forms of punishment) overshadow social preferences. But this in no way implies that these preferences do not exist, nor that they played no role in the evolution of human sociality.

A comparative approach can be helpful in illuminating just how important social preferences are. Guala rightly advocates a comparative anthropological approach. Extending this approach to animals is drawing attention to just how special social preferences might be. Animals do behave altruistically and punitively in the wild (and in captivity). Or, at least, they appear to. For instance, they share food with others (de Waal et al. 1993) and punish non-cooperators (Hauser 1992). However, the benefits are likely immediate rather than delayed, casting social preferences in doubt: Food sharing, at least in chimpanzees, can be explained by surrendering to begging pressure (Gilby 2006), and punishment of failure to give food calls is likely due to conflicts over contested food (Gros-Louis 2004).

The approaches taken by experimental economists are being adapted for other animals. An adaptation of the dictator game has one animal choose between one of two food trays, resulting in mutually beneficial outcomes (1/1) as opposed to purely selfish ones (1/0), or altruistic outcomes (0/1 vs. 0/0). Chimpanzees do not show anything resembling a prosocial preference in these studies. Results for cooperatively breeding primates such as cottontop tamarins are mixed (for reviews, see Silk & House 2011; Jensen, in press). It may be that the competitive nature of chimpanzees in feeding contexts interferes with prosociality. Yet, when given a choice between an altruistic outcome or a spiteful one, chimpanzees remain indifferent (Jensen et al. 2006). Furthermore, while they will retaliate against harmful actions – namely, theft of their food – they do not respond spitefully to unfair outcomes (another chimpanzee eating food taken from the subject by the experimenter; Jensen et al. 2007a). The latter study is similar in spirit to a money-burning game in which people will forfeit money (in the lab) to see someone else suffer a cost (Zizzo & Oswald 2001).

The Ultimatum Game, as described by Guala, is a widely used tool for testing social preferences in response to unfairness. A reduced form of the Ultimatum Game pits rejections of unfair outcomes (80/20) against alternatives that are either fair (50/50), generous (hyper-fair: 20/80), no different (80/20) and very unfair (hyper-unfair: 100/0) (Falk et al. 2003). The results in the lab suggest that rejections are due to a sensitivity to both unfair outcomes and unfair intentions (Falk & Fischbacher 2006). When chimpanzees were confronted with a similar dilemma, proposers did not choose fair outcomes and responders accepted all nonzero offers (Jensen et al. 2007b). Chimpanzees, at least, appear to behave like rational maximizers in the sense that they are indifferent to unfair outcomes. Social preferences do not seem to play a role in punitive and spiteful behaviours in our closest living relatives (Jensen 2010; Jensen & Tomasello 2010).

Third-party punishment is only fleetingly mentioned by Guala, despite its possible importance in the real world. Self-interested (second-party) punishment, rather than altruistic punishment, may be important in small-scale societies (Marlowe & Berbesque 2008). Yet third-party punishment does appear to play a role in maintaining cooperation, even in the absence of institutionalised or collective punishment (Henrich et al. 2006). Social preferences that govern punishment of behaviours that affect us personally can extend to other individuals, and from this, allow for the emergence of rules that ought to apply to others (norms). How these sensitivities evolved is poorly understood. Looking for third-party punishment in other animals in experimental contexts will inform the debate on the role of social preferences in our sociality and help us decide whether strong reciprocity can be widely read.

Results from experiments on animals provide a contrast to the results of experiments on humans. Concern for outcomes affecting others does play a role in guiding social choices for humans – and possibly only humans – in the lab. It would be surprising if such robust experimental findings whither in the light of the real world.

The strategic logic of costly punishment necessitates natural field experiments, and at least one such experiment exists

doi:10.1017/S0140525X11000926

Tim Johnson

Department of Political Science, Stanford University, Stanford, CA 94305; Atkinson Graduate School of Management, Willamette University, Salem, OR 97301.

timj@stanford.edu www.stanford.edu/~timj

Abstract: Costly punishment's scarcity "in the wild" does not belie strong reciprocity theory as Guala claims. In the presence of strong reciprocators, strategic defectors will cooperate and sanctioning will not occur. Accordingly, natural field experiments are necessary to assess a "wide" reading of costly punishment experiments. One such field experiment exists, and it supports the hypothesis that costly punishment promotes cooperation.

Although he persuasively argues that field evidence is needed to support a "wide" reading of costly punishment experiments, Guala errs in his assessment of the current state of non-laboratory evidence concerning strong reciprocity. First, Guala holds that the dearth of punishment-induced cooperation in the anthropological record undermines a wide reading of costly punishment experiments (target article, Abstract). Second, Guala holds that "there are no natural field experiments on costly punishment" (sect. 7, para. 5). Both of these claims are problematic. Guala's first claim fails to appreciate the strategic logic of costly punishment, which predicts that punishment will occur rarely if it proves effective at fostering cooperation. Guala's second claim results from an incomplete reading of the past literature,

which, as I show here, contains at least one natural field experiment concerning costly punishment. Once these problems with Guala's argument are addressed, it becomes clear that field evidence supports a wide reading of costly punishment experiments.

Guala's first claim – which holds that "there is no evidence that cooperation in the small egalitarian societies studied by anthropologists is enforced by means of costly punishment" (Abstract) – seems especially damning to a wide reading of costly punishment experiments. If costly punishment does not exist outside the laboratory, then it cannot support cooperation in the everyday world. Albeit intuitive, that reasoning fails to consider the strategic logic of costly punishment. If punishment makes free-riding more expensive than cooperation, and if a population contains myriad strong reciprocators willing to punish free-riders, then strategic defectors will opt to cooperate and strong reciprocators will have no need to engage in costly punishment. In such equilibrium, observers would not witness any *actual* costly punishment: cooperation induced by costly punishment and cooperation produced by other mechanisms would appear "observationally equivalent" (for a general discussion, see Weingast & Moran1983, p. 767, fn. 2). Guala ignores this implication of costly punishment's strategic logic, and in so doing, he fails to realize that the absence of costly punishment from the anthropological record remains consistent with a wide reading of costly punishment experiments.

Also, by ignoring the strategic logic of costly punishment, Guala's point about the importance of natural field experiments carries less force than it should (sect. 6). The strategic logic of costly punishment indicates that true natural field experiments are crucial to test strong reciprocity theory outside the laboratory. Without exogenous variation in either (1) the opportunity to engage in costly punishment (which would allow defectors to demonstrate that they will free-ride if sanctioning is impossible) or (2) the occurrence of free-riding (which would allow strong reciprocators to demonstrate their willingness to punish), it is impossible to assess whether cooperation outside the laboratory results from costly punishment or some other mechanism. In light of this need for experiments, Guala casts a threatening shadow over strong reciprocity theory by claiming that "there are no natural field experiments on costly punishment" (sect. 7, para. 5).

Guala's claim, however, is incorrect. At least one natural field experiment provides evidence concerning the influence of costly punishment on cooperation. In a study of voter turnout, Gerber et al. (2008) told a randomly selected group of citizens that experimenters would send information about their voter turnout record – along with their names and addresses – to other citizens in the aftermath of a forthcoming election. The revelation of this information raised the possibility that fellow citizens could identify and sanction individuals who did not engage in personally costly – yet group benefiting – turnout. In so doing, the experimental manipulation exogenously influenced individuals' awareness of punishment possibilities, thus creating conditions in which experimenters could examine whether the introduction of costly sanctioning opportunities would, in fact, increase cooperation. Indeed, consistent with laboratory evidence showing immediate increases in cooperation when experimenters introduce the opportunity to punish (see Fehr & Gächter 2000a, p. 986, Figs. 1A and 1B; Fehr & Gächter 2002, p. 138, Figs. 2A and 2B), the mere possibility of costly punishment increased cooperative voter turnout by roughly 8 percentage points, which corresponded to an approximate 27% change from the control group's turnout rate (Gerber et al. 2008). Although it was designed to illuminate the mechanisms underlying electoral participation, the field experiment conducted by Gerber et al. (2008) bears directly on costly punishment and it indicates that uncoordinated costly punishment *can* increase cooperation outside the lab.

Other studies that combine observational and experimental data provide complementary evidence, while also showing an additional means to test the plausibility of laboratory findings concerning costly punishment. For instance, Smirnov et al. (2010) show that individuals who incur the costs of cooperation and punishment in laboratory public goods games exhibit a greater likelihood of engaging in partisan collective action in their daily lives. Given that maintaining political parties represents a prototypical public goods problem (Aldrich 1995, p. 31), these findings imply that organizations engaged in non-laboratory cooperative enterprises may succeed – at least in some significant part – precisely because they are populated by strong reciprocators (Smirnov et al. 2010). Not only do those findings offer insight into non-laboratory cooperation, but they also illustrate another means by which researchers can examine whether behavior observed in laboratory costly punishment experiments corresponds with conduct outside the lab.

Ultimately, such concerns about the external validity of costly punishment experiments are important and Guala deserves credit for voicing them. Yet, in the end, those concerns are less pressing than Guala claims. Contrary to his suggestions, scientists should not expect the anthropological record to contain examples of costly punishment; the strategic logic of costly punishment holds that no such examples will exist if punishment truly deters free-riding. Nor should scientists doubt a wide reading of costly punishment experiments due to a lack of natural field experiments; at least one such experiment exists and it supports the hypothesis that costly punishment facilitates cooperation.

Altruistic punishment: What field data can (and cannot) demonstrate

doi:10.1017/S0140525X11001270

Nikos Nikiforakis

Department of Economics, University of Melbourne, 3010 Victoria, Australia.
n.nikiforakis@unimelb.edu.au
http://www.economics.unimelb.edu.au/staff/nikosn/

Abstract: The rarity of altruistic punishment in small-scale societies should not be interpreted as evidence that altruistic punishment is not an important determinant of cooperation in general. While it is essential to collect field data on altruistic punishment, this kind of data has limitations. Laboratory experiments can help shed light on the role of altruistic punishment "in the wild."

Laboratory experiments have provided evidence that many individuals are willing to punish at a personal cost those favoring their private over the public interest. This type of punishment has been dubbed "altruistic" because it benefits third parties by discouraging free-riding (Fehr & Gächter 2002). Guala's article is an excellent critical review of the literature on altruistic punishment.[1] The careful discussion of the field data on altruistic punishment is particularly useful. It is important to note, however, that this data comes from small-scale societies. While altruistic punishment could, in principle, explain cooperation when individuals interact repeatedly, one should not hurry to infer by the infrequency (or even the absence) of altruistic punishment in these cases that it is not an important force in supporting cooperation in general. The reason is that the expected cost of altruistic punishment can be larger in repeated than in one-shot interactions. For example, free-riders have an incentive to counter-punish in repeated interactions to signal that punishment will not be tolerated in the future.

Counter-punishment raises the cost of altruistic punishment and can also lead to feuds (i.e., cycles of retaliation) that further increase the cost of enforcing cooperation. Altruistic punishment can also destroy social ties, which means that future

benefits from interacting with a particular individual are foregone. The demand for altruistic punishment has been shown to decline when the cost of punishment increases, all else equal, in laboratory experiments (Anderson & Putterman 2006; Carpenter 2007; Egas & Riedl 2008; Nikiforakis & Normann 2008). Therefore, the rarity of altruistic punishment in small-scale societies does not necessarily imply that altruistic punishment plays no role in one-shot interactions. In addition, recent studies have found that individuals can be quite forward looking if there exists a prospect of future interactions (Cabral et al. 2011; Reuben & Suetens 2011).

The existence of altruistic punishment can be tested more clearly in one-shot interactions. To my knowledge, the only experimental field evidence on altruistic punishment in one-shot interactions comes from a natural field experiment recently run by Balafoutas and Nikiforakis (2011) in Athens, Greece. In this study, the authors exogenously violated two well-established, efficiency-enhancing social norms (littering and standing on the left side of escalators in a central subway station). The goal was to examine whether civilians punish norm violators. The individuals were unaware they were taking part in an experiment. The rate of altruistic punishment is overall low: Altruistic punishment is observed in only 11.7% of violations (35 cases out of 300). Surprisingly, violations of the more well-established of the two norms (littering) are less likely to be punished (4% versus 19.3%). Questionnaire data indicates that the reason for the low occurrence of altruistic punishment is that people are concerned about being counter-punished by the norm violator. Further, they consider individuals who litter more likely to counter-punish than those who stand on the left side of the escalators. Interestingly, the vast majority of people do adhere to the two norms. This raises the question whether the low frequency of altruistic punishment in our experiment can explain by itself the widespread adherence to the norm that is observed. We believe that this is unlikely to be the case. For example, some individuals will be unwilling to litter because they have internalized this norm early in their lives. However, one cannot rule out that altruistic punishment plays a significant disciplining role. It is possible that some individuals may not litter because they are concerned that they will be punished: The probability may be low, but it is not zero.

Is there any evidence that the threat of altruistic punishment can sustain cooperation even when the overall threat of punishment is very low? Nikiforakis and Engelmann (2011) study a public-good game in which altruistic punishment could lead to counter-punishment and lengthy feuds. The design imposes minimal restrictions on the punishment strategies subjects can adopt and thus allows them to use a range of complex strategies that are often found in the field (e.g., punish non-punishers, intervene to stop a feud). The most surprising finding is that the likelihood that an extreme free-rider gets punished is as low as that of a cooperator being punished. However, despite the low frequency and severity of altruistic punishment, cooperation rates are higher than those typically observed in public-good experiments without any punishment opportunities. This suggests that the mere possibility of triggering a feud may be sufficient to stop some individuals from free-riding. Therefore, the results in Nikiforakis and Engelmann (2011) suggest that the low frequency of altruistic punishment reported in the field experiment of Balafoutas and Nikiforakis (2011) may be sufficient to support (at least to some extent) cooperation.

I am referring to the study by Nikiforakis and Engelmann (2011) to emphasize that laboratory experiments can help us understand the determinants of cooperation in the field by allowing researchers to study counterfactuals (e.g., cooperation in the absence of altruistic punishment) in ways that is difficult (if not impossible) to do in the field. Field data – both experimental and not – is essential in order to understand the forces that exist outside the laboratory, the strategies employed by individuals, and the different institutions that emerge in different circumstances. However, field data will prove insufficient in some cases to explain the determinants of cooperation by itself. The reason is that key variables such as the perceived threat posed by altruistic punishment and the risk preferences of potential norm violators are difficult to measure in the field. The difficulty of measuring important variables in the field and the limited control of researchers over the experimental environment (e.g., researchers may not be able to remove a certain strategy in a natural field experiment) is what makes laboratory experiments useful in understanding factors that facilitate cooperation such as altruistic punishment.

NOTE

1. Guala uses the term "costly punishment" instead of altruistic punishment, but since the article focuses on the impact of punishment on cooperation, altruistic and costly punishment refer to the same type of behavior.

Experiments combining communication with punishment options demonstrate how individuals can overcome social dilemmas

doi:10.1017/S0140525X11001282

Elinor Ostrom

Workshop in Political Theory and Policy Analysis, Indiana University, Bloomington, IN 47408.

ostrom@indiana.edu http://www.indiana.edu/~workshop/

Abstract: Guala raises important questions about the misinterpretation of experimental studies that have found that subjects engage in costly punishment. Instead of positing that punishment is the solution for social dilemmas, earlier research posited that when individuals facing a social dilemma agreed on their own rules and used graduated sanctions, they were more likely to have robust solutions over time.

As a social scientist who conducts extensive field research, as well as doing experiments and theoretical work, I find the target article resonates well with my own research. Since one of the earliest experiments on punishment was conducted by myself, Roy Gardner, and James Walker (Ostrom et al. 1992), the origin of that article may be interesting to readers of this issue. As Guala indicates, we had earlier reviewed a very large number of in-depth case studies of settings where users organized their own governance system related to common-pool resources. In my 1990 book, I reported on a massive effort to synthesize findings from this large number of studies. One of the principles that I derived from this study was that long-lasting and robust institutions tended to use "graduated sanctions" (Ostrom 1990). By graduated, I meant that users of successful common-property regimes would sanction one another for observed nonconformance to their own rules and would first gently remind one another of such infractions. The gentle reminders would be exercised one or two times, but after using "shame" to try to bring someone around to following rules, there would be other punishments that would be imposed in an ever-greater level of cost to the recipient. The final punishment might be rather severe.

I also observed punishments being administered in the field and was puzzled, because one could not explain such punishments by users themselves using game theory. Therefore, I asked Roy Gardner to develop a rigorous game-theoretic model and worked with James Walker to test that model in the lab.

When we gave the participants in an experiment an opportunity to pay a fee to fine someone else, they did indeed use it contra to the game theory prediction. In fact, we found they overused it. We did not quite know how to explain the overuse, but we finally used the term "blind revenge" to explain the fact that frequently the sanctioner would direct the fines against others whose computerized record showed that they were highly

cooperative. We thought they probably figured that the cooperators were initially fining the non-cooperators and the non-cooperators then fined the cooperators in revenge.

In the massive case studies we had worked through, we did not find many cases of blind revenge. Hence, we decided to move to the next step in the lab and give participants an opportunity to communicate and decide on their own rules. Those who engaged in self-governance then did not use fines very often. They increased their levels of cooperation to the point that their net benefits were very close to optimal. Thus, the combination of agreement and discussion first and evidence that others were cooperating, led to a much better result than an externally designed sanctioning system without a set of rules that the participants had agreed to.

Although in the Janssen et al. (2010) study we did not give participants an opportunity to design their own rules, we gave them the opportunity to simultaneously engage in communication and use of fines, the use of fines alone, and the use of communication alone. As Guala comments, when communication and punishment opportunities were combined, the subjects did the very best in the lab.

There are several issues being debated by a variety of very distinguished scholars. There is no question that humans have the capability of engaging in serious punishment of each other; but that should not lead us to conclude that the way of achieving long-term sustainability is by enabling participants to punish each other without enabling them to engage in serious discourse about the rules they want to adopt and how they should be observed and sanctioned. When participants in a dilemma setting are able to engage in serious discussion, debate about their joint future, and agree on rules that limit strategies, they have much less need to use punishment against defectors. Monitoring each other and initially shaming those who do not comply with their rules is, however, an essential component for sustaining that cooperation over time. Stronger sanctions are not often needed, but their authorization backs up the use of mild sanctions when rule-breaking behavior is initially observed. Our recent research related to forestry institutions around the world demonstrates that when the users monitor each other's behavior in a forest, forest conditions are substantially enhanced (Coleman 2009; Coleman & Steed 2009; Chhatre & Agrawal 2009).

I am glad to see these issues being raised in a way that makes it possible to move forward to a better understanding of the role of punishment in overcoming social dilemmas of various kinds.

Importing social preferences across contexts and the pitfall of over-generalization across theories

doi:10.1017/S0140525X11001294

Anne C. Pisor[a] and Daniel M. T. Fessler[b]

[a]*Department of Anthropology, University of California, Santa Barbara, CA 93106-3210;* [b]*Department of Anthropology and Center for Behavior, Evolution, and Culture, University of California, Los Angeles, Los Angeles, CA 90095-1553.*

pisor@umail.ucsb.edu **dfessler@anthro.ucla.edu**

http://www.uweb.ucsb.edu/~pisor/

http://www.sscnet.ucla.edu/anthro/faculty/fessler

Abstract: Claims regarding negative strong reciprocity do indeed rest on experiments lacking established external validity, often without even a small "menu of options." Guala's review should prompt strong reciprocity proponents to extend the real-world validity of their work, exploring the preferences participants bring to experiments. That said, Guala's approach fails to differentiate among group selection approaches and glosses over cross-cultural variability.

We agree with Guala that it can be difficult to draw conclusions about human evolution from highly controlled experimental games. Controlling any and all third variables facilitates replication and repetition, enabling comparison of behavior across experiments (Guala 2005). However, striving for internal validity introduces a double-edged sword: Economic games provide insight, but they present only a rough approximation of the real world. Strong reciprocity arguments often strive to connect game play to real life by citing anecdotal evidence. Nevertheless, though we endorse caution in interpreting experiments, Guala himself overlooks the incorporation of real-world aspects into recent field-based economic games. This research allows greater insight into the societies under investigation. Moreover, we take issue with both Guala's homogenizing account of group selection theories and his failure to acknowledge variability across subsistence groups. That said, we believe the present article should spur strong reciprocity theorists to further explore the variable social preferences exhibited by participants.

A narrow interpretation of experimental economic games – an uncontroversial reading of the evidence, as Guala notes – suggests that "punishment mechanisms are useful *methodological devices to observe social preferences*" (sect. 5, para. 2, italics in original). We agree. These social preferences have been sometimes termed "informal norms," including norms of fairness and reciprocity (Guala 2008). For example, Western participants are annoyed and often angry if another participant has a larger net gain than they do (Dawes et al. 2007; Fehr & Gächter 2002). If internal validity is rigorously sought within experiments and maintained across experiments via replication, we can expect any differences in game play to correspond to differences in social preferences applied by participants to the experimental context. The issue with external validity arises because game play, by virtue of experimental control, is far-removed from the real-life situations strong reciprocity theorists seek to explain.

We agree that the simple design of many games necessitates caution in interpretation. We do not mean that all economic games are overly simplistic, but that it is difficult to make inferences without control groups of sorts. Many experiments cited by strong reciprocity theorists do not allow for coalition formation, reputation building, or less expensive punishment options. Rockenbach and Milinski (2006) found that reputation formation matters: Public goods contributions were greater when costly punishment and indirect reciprocity (i.e., withholding cooperation) were united with reputation building, and participants preferred to join these groups. Similarly, Jacquet et al. (2011) have demonstrated that both negative and positive reputational consequences external to the game context enhance cooperation within the game context. Egas and Riedl (2008) found that low-cost and high-impact punishment best promotes cooperation. These results support the idea that costly punishment is probably not as common when "the full menu of strategies" (target article, sect. 6, para. 3) is available.

Guala's primary concern is the extent to which economic games reflect punishment mechanisms "in the wild." Though in the past Guala (2008) applauded the MacArthur Foundation–sponsored Economic Man studies, in the present article he emphasizes that the incidental introduction of cultural practice by some researchers (Henrich et al. 2005, Table 4) and participants (Henrich et al. 2005, sect. 8) is not equivalent to an experiment designed to reflect on the particular population of study. Guala is not the only observer to raise concerns about the external validity of field-based games such as Economic Man (e.g., Gurven & Winking 2008); however, he overlooks more recent efforts to bring external validity to economic games in the field context. Notable recent field studies have endeavored to match games to context, and to derive clear insights about costly monitoring and punishment within a particular cultural group (see Jack 2009; Lamba & Mace 2010; Rustagi et al. 2010).

Despite the strengths of the present review, Guala risks the same pitfall for which he criticizes others: over-generalization.

His failure to differentiate among theories of (what he terms) "group selection" does a disservice to the understanding of this area of study. There is a significant difference between biological group selection and gene-culture coevolution (for discussion, see West et al. 2011). By describing propensities to internalize norms as an aspect of our innate psychology and explaining the cooperation-enhancement of some norms as the product of cultural group selection, gene-culture coevolution theory affords greater variability across groups than does biological group selection theory. While important, these distinctions are admittedly sometimes obscured in the literature, even though connections can be separately drawn between biological group selection and strong reciprocity, and between gene-culture coevolution and strong reciprocity (e.g., Fehr et al. 2002).

The above distinctions are important because, at the empirical level, Guala provides a simplified view of small-scale societies that minimizes variation among them. Guala distills the variety of punishment behaviors outlined in Boehm's (1999) research, drawing generalizations about homicide in hunter-gatherers, among other things. Boehm (1999) himself reports that a dearth of punishment data required him to use unsystematic methods of sampling. Today, better archived ethnographic materials afford more systematic gleaning of examples of punishment (though the cases themselves remain anecdotal). Additionally, by dwelling on fission-fusion as a conflict management strategy, Guala overlooks ecological variation that influences the availability of this strategy. For example, 25% of hunter-gatherers in a sample of 340 societies are actually sedentary (Marlowe 2005), making fission a less ready solution for conflict.

Guala's review of negative strong reciprocity provides a useful platform for subsequent work. We would like to see more even-handed treatment of both the relevant theories and the available ethnographic data. That said, we agree that, regardless of their cultural group, participants face a contrived social situation in economic game experiments. Investigators need to focus on the preferences participants bring to experimental games, including (1) explaining the origins of these preferences, (2) understanding how they manifest in real-world situations, and (3) accounting for individual- and group-level differences in preferences.

Culture: The missing piece in theories of weak and strong reciprocity

doi:10.1017/S0140525X11001300

Dwight Read

Department of Anthropology, University of California, Los Angeles, Los Angeles, CA 90095.

dread@anthro.ucla.edu

http://www.anthro.ucla.edu/people/faculty?lid=886

Abstract: Guala does not go far enough in his critique of the assumption that human decisions about sharing made in the context of experimental game conditions accurately reflect decision-making under real conditions. Sharing of hunted animals is constrained by cultural rules and is not "spontaneous cooperation" as assumed in models of weak and strong reciprocity. Missing in these models is the cultural basis of sharing that makes it a group property rather than an individual one.

Guala rightly draws attention to the fact that human decision-making under experimental game conditions cannot be extrapolated directly to decisions made under the real conditions. For example, the Ju/'hoansi (!Kung san) make decisions in real life that contradict their behavior in experimental game conditions (Wiessner 2009). The disconnect relates to decisions being predicated on both a biological and cultural heritage (Read 2010a); hence, the behaviors observed in a game context are a complex mixture of background predispositions and the conditions specified in the game context, and need not mirror decisions made during daily life.

Guala does not go far enough, though, in his discussion of the disjunction between experimental and real conditions. In an endnote he observes that current theories of reciprocity based on game theory have not drawn upon the concept of reciprocity previously developed in anthropology (e.g., Sahlins 1972/1974) to account for the informal exchange of goods and services that is part of social life in human societies (see Note 2 in the target article). However, he does not follow up on his observation and instead limits his argument to a discourse on weak versus strong reciprocity, as if the only matter at issue is whether we account for cooperative behavior in human societies by one or the other of these two competing theories.

Running deeper than the surface issue of whether experimental evidence for strong reciprocity can be extrapolated to behaviors in natural conditions, is whether our perception of cooperative behavior in human societies has been framed correctly in the first place. Guala, like most researchers in this area, accepts uncritically the notion that small-scale human societies such as hunter-gatherers can be characterized as "acephalous social orders based on *spontaneous cooperation*" (sect. 8, para. 2, emphasis added). From the assumption of "spontaneous cooperation," it follows that relevant evolutionary questions are: How and under what conditions will there be selection at the individual level for cooperation as a trait? And, how and under what conditions will a population composed of individuals with a cooperation trait be stable against invasion by a "free-rider" trait?

The problem with formulating cooperation in this manner, along with its attendant questions, lies in the lack of evidence that individuals in small-scale societies are spontaneous cooperators. Consider how food resources are shared. In a hunter-gatherer society such as the Ju/'hoansi, food resources "in the wild" are collectively owned, in the sense of collective rights of access, by a residence group of families closely related through culturally constituted kinship relations (see Read 2001; 2007, for the difference between biological and cultural kinship). Rights to those food resources depend on membership (temporary or permanent) in a residence group. Collective ownership of resources changes into individual ownership by the mode of procurement and characteristics of the resources. Resources that come in small units are accessible to any able-bodied adult on a regular basis, and have low risk of failure on each procurement episode (such as, but not limited to, vegetal foods); these are transformed from collective into individual ownership through foraging. Resources that come in relatively large units are differentially accessible by adults according to individual skills, and have high risk of failure on a given procurement episode (such as game animals); these resources are considered to be collectively owned through hunting. For the latter, ownership changes from collective to individual according to culturally specified rules of sharing that remove decisions about sharing from the individual hunter to the group as a collectivity. Among the Ju/'hoansi, for example, the cultural rule is that the owner of the arrow that killed the animal (who need not have been present during the hunt) distributes the meat from the animal (Marshall 1976). Among the Netsilik Inuit, seals killed in winter hunting through their breathing holes in the pack ice were distributed in accordance with a culturally constituted system of "sealing partners" (Balikci 1970). Cultural rules like this make meat-sharing a group-level, not an individual-level, trait (Read 2012).

In general, resources that are individually owned are not subject to cultural rules of sharing. Individually owned resources are shared within a family (with a culture specific definition of what constitutes a family) and without cultural rules. Sharing within a family corresponds to "spontaneous cooperation." However, we need neither weak nor strong reciprocity to account for sharing and cooperation within a family.

Individually owned resources allow for individual decisions about whether they should be kept or given to others, but the

act of gift giving is a social one (Mauss 1924/1990). Gift-giving is subject to cultural rules such as generalized reciprocity (Sahlins 1972/1974) in which A gives to B, with the (usually unstated) understanding that B will reciprocate at some indefinite time in the future and by an unspecified amount, as in *hxaro* gift giving among the Ju/'hoansi (Wiessner 1977; 1982). Generalized reciprocity is dependent upon trust by the parties concerned (Sahlins 1972/1974), and trust depends on close kin relations (kin being understood in a cultural and not a biological sense). This is the context where "punishment" comes into play, but punishment, as Guala discusses, is not of the kind invoked in the theory of strong reciprocity. Rather, it is social punishment in which the transgressor is made to understand by various means that her or his behavior is unacceptable as a kinsman. This, as pointed out by Guala, is punishment by the collective against the individual, by one's kin against oneself. Among close kinsmen, social punishment is effective because of one's dependency on kin for surviving in hunter-gatherer and other small-scale societies, not because of the magnitude of the punishment in a material sense.

Cooperative behavior characterizes sharing of collectively owned resources, but it is not "spontaneous cooperation" and instead is determined through cultural rules. The specificity of the cultural rules relates to the degree of risk that failure to succeed in a resource procurement episode has for the survival of members of the group. For the Netsilik Inuit, disputes over sharing of seals hunted through the pack ice in winter could lead to the breakdown of a winter sealing camp, thereby exposing camp members to the risk of death through starvation (Balikci 1970). Correspondingly, the Netsilik had extensive and highly specific rules for the sharing of seals that transformed the individual hunter into an agent for the group and removed decisions about sharing from the hunter. In effect, the cost of a dispute over sharing of seals was too high and too immediate to allow for the sharing to be subject to individual decision-making even if there were costly punishment as hypothesized under strong reciprocity. In contrast, the Tiwi living on Melville and Bathurst Islands off the northern coast of Australia had low risk and relatively simple rules for sharing of hunter meat (Goodale 1971). (Risk can be measured indirectly by the complexity of implements used in resource procurement [Read 2008; Torrence 1989]. The Netsilik used complex implements, and the Tiwi simple ones.)

The notion of cooperative behavior used in weak or strong reciprocity theories is at odds with the facts of meat-sharing in hunter-gatherer societies. These theories do not take into account the major transformation that took place in the basis for social organization during the evolution of human societies (Read 2010b; 2012). That transformation is from societies in which patterns of social organization and structure emerge from face-to-face interaction of group members to relation-based societies (Read 2012) predicated on behaviors formed in accordance with systems of organization for the society as a whole (Leaf & Read, in press), such as culturally constructed systems of kinship relations that define the boundary for, and internal organization of, the small-scale human societies from which are derived all larger-scale human societies.

Towards a unified theory of reciprocity

doi:10.1017/S0140525X11001312

Alejandro Rosas

Department of Philosophy, Universidad Nacional de Colombia, Bogotá, Colombia, and Konrad Lorenz Institute (KLI) for Evolution and Cognition Research, 3422 Altenberg, Austria.

arosasl@unal.edu.co

http://www.docentes.unal.edu.co/arosasl/

Abstract: In a unified theory of human reciprocity, the strong and weak forms are similar because neither is biologically altruistic and both require normative motivation to support cooperation. However, strong reciprocity is necessary to support cooperation in public goods games. It involves inflicting costs on defectors; and though the costs for punishers are recouped, recouping costs requires complex institutions that would not have emerged if weak reciprocity had been enough.

Guala engagingly criticizes the claim that strong reciprocity (SR) explains human cooperation beyond the limitations of classic or weak forms of reciprocity (WR) (direct and indirect reciprocity). For advocates of SR, its unique success comes from two distinctive characteristics: It is biologically altruistic, and it involves normative motivation. Guala shows convincingly that there is no evidence of a biologically altruistic form of human reciprocity outside the lab. I shall argue, furthermore, that normative motivation is present in weak reciprocity as well. In the end, SR is more similar to weak reciprocity than its advocates would admit. Nonetheless, there is some novelty to SR, although of a more modest nature.

SR has four properties as initially introduced: (1) In contrast to withdrawing cooperation, SR entails a net cost for the strong reciprocator, both in the short and the long term and therefore evolves through group selection; (2) it inflicts a direct cost on defectors; (3) repeated Public Goods games (PGGs; i.e., *n*-person Prisoner's Dilemmas [PDs]) are its proper domain; (4) it involves norms as socio-psychological mechanisms. Of these four properties, Guala shows that the first one is an artefact of lab experiments, with no real counterpart in the field. However, the other three properties are instantiated often enough in human institutions.

Properties (2) and (3) alone justify a distinction between SR and WR. In PGGs (repeated *n*-person PDs where $n \gg 2$) cooperators cannot discipline defectors by withdrawing contributions in subsequent periods, as in direct and indirect reciprocity (Fehr & Gächter 2000a; Ledyard 1995). Therefore, a second stage of the game is required, designed as a two-person game, where cooperators target and inflict costs on individual defectors. This two-stage structure is common to both the lab and the field. However, punishers in the lab pay a fee, usually less than the cost inflicted on the punished. Because retaliation against punishers is possible in the field, there is some reason to view this as unrealistic. But far more common in the real world is that humans device institutions to compensate punishers for their costs. Tax officials fining tax evaders have usually little to fear from them in terms of retaliation. Beyond that, they are paid for their efforts. Guala's convincing argument is that most field examples of SR are like this case. The property listed above as (1) has to be dropped, even if it was originally crucial.

Guala's view is that WR alone is at work in the field whenever punishers are compensated for inflicting a direct cost on defectors, because no net costs are involved in the long term. But the fact remains that in PGGs cooperation is enforced in a second stage where punishers incur short-term costs that have to be institutionally compensated. In some cases it is possible to design a second stage where cooperators withdraw cooperation from those who defected in the PGG (Milinski et al. 2002). Defectors experience this as punishment and increase their contributions. But this design is not always possible: Tax evaders have to be fined. There is, therefore, a deep theoretical motivation to introduce a form of reciprocity that differs from direct and indirect reciprocity. It is required in relation to PGGs and leads directly to complex state or state-like institutions like the *Carte di Regola*. If defectors in PGGs could be punished by withdrawing benefits from them in direct or indirect reciprocity games, humans would not have invented those institutions.

Almost every worker in this field sees reciprocal altruists as "solely concerned about future gains" (target article, sect. 3, para. 3). In contrast, "Strong reciprocators cooperate because they feel it is the right thing to do, and they are ready to punish defectors at a cost" (sect. 3, para. 3). I want to argue, instead, that moral normativity is present in every form of

human reciprocity. This claim can be supported historically, simply by showing that Trivers had fairness in mind when he speculated about the psychological mechanism that supports reciprocal altruism. Trivers' reciprocal altruists demand genuine psychological altruism from their partners: "Individuals who initiate altruistic acts out of a calculating rather than a generous-hearted disposition" (Trivers 1971, p. 51) are "subtle cheaters." Reciprocal altruists demand partners that signal character traits like fairness. Evolutionary game theory designs strategies that are more generous or forgiving than Tit for Tat, and also more successful in the evolutionary dynamics (Kollock 1993). In the real world, these behavior patterns signal unselfish characters, exactly what human players are looking for in the field.

Alternatively, we can reflect about the limitations of WR as pointed out by advocates of SR. If one makes too close an analogy between WR and the "game-theoretic result known as the *folk theorem*" (target article, sect. 2, para. 7, italics in original), advocates of SR immediately object that, in this case, WR is only applicable when the conditions of the folk theorem hold. In fact, these conditions never hold. But it seems inconsistent for critics of WR to point this out and nonetheless insist that Trivers' reciprocal altruism is the instantiation of pure self-interest leading to cooperation in the – nonexistent – conditions of the folk theorem. Humans display reciprocal altruism only because of a special psychology that substitutes for the absence of those conditions. For example, as Frank (1988) has shown, reciprocal altruism would not be possible without the moral emotions counteracting the effects of temporal discounting. Moreover, as shown in simulations of indirect reciprocity, one-shot encounters in large groups are not an obstacle to cooperation because humans have a real concern for reputation, which circulates in a group through gossip. And despite the lack of complete information regarding reputations, humans extrapolate the bits of information they can get hold of to character traits stored in semantic memory. The emotions that solve the discounting problem, and the selection of prospective partners according to character traits like fairness, make reciprocators in dyadic games *psychologically* unselfish, though they remain *biologically* selfish by avoiding the costs of mutual defection. Note that direct reciprocity in PDs is rare in nature; it requires special cognitive and psychological traits that few organisms besides humans can display.

Special human vulnerability to low-cost collective punishment

doi:10.1017/S0140525X11000896

Don Ross

University of Cape Town, School of Economics, University of Cape Town, Private bag, Rondebosch, 7701 Cape Town, South Africa.

don.ross@uct.ac.za http://uct.academia.edu/DonRoss

Abstract: Guala notes that low-cost punishment is the main mechanism that deters free-riding in small human communities. This mechanism is complemented by unusual human vulnerability to gossip. Defenders of an evolutionary discontinuity supporting human sociality might seize on this as an alternative to enjoyment of moralistic aggression as a special adaptation. However, the more basic adaptation of language likely suffices.

Guala performs an invaluable service in clarifying the absence of convincing empirical evidence for strong reciprocity as a likely mechanism in the evolution and maintenance of modern human sociality. As his survey shows, humans in small groups without government generally defend their institutions and cooperative norms against free-riding by the use of very low-cost sanctions that are applied collectively. This implies that Occam's razor should be wielded against the proposals of some theorists that hominid evolution involved an evolutionary discontinuity that produced emotional satisfaction from moralistic punishment sufficient in motivational strength to subvert Hamilton's rule. However, there is another side to the picture that Guala does not address, and which could yet motivate supporters of evolutionary discontinuity hypotheses. The effectiveness of low-cost social punishment among humans should lead us to ask whether they are unusually vulnerable to such sanctions, and, if so, whether a genetic change after the divergence of hominids from the main ape branch might have directly promoted such vulnerability.

As Guala reports, ethnographic evidence suggests that the crucial condition for the effectiveness of socially distributed low-cost punishment is the high aversiveness of emotions associated with shame and social guilt. Creative literature and reflective memoirs from all literate societies abound with accounts of characters driven to extreme behavior, especially suicide and homicide, by psychic discomfort associated with experienced shame. The psychiatric clinical literature supports these culturally familiar interpretations (Lewis 1995). For an organism susceptible to such severe emotional pain from the knowledge that, in the judgment of fellow community members, he or she has diverged from prevailing normative expectations, it is rational to take great care to minimize the probability of being an object of negative gossip. Recognition of this point leads naturally to the following hypothesis: Perhaps the crucial device for controlling free-riding in humans is an evolved disposition to suffer severely from awareness that one is widely perceived as normatively deviant. To the extent that this hypothesis is deemed worthy of scientific attention, we will then be led to wonder whether the disposition in question is a direct product of genetic selection and, if so, whether it is specific to humans.

The question of the generality of shame responses in social animals is so far neglected. De Waal (1996) interprets some behavior of subordinate monkeys after indulging in copulations that would be punished by dominants if they were present, as indicating precursors of shame and/or social guilt. However, a recent extensive survey of cognitive structures and mechanisms that support coordination and cooperation in intelligent social animals (Emery et al. 2007) contains no index entries for these emotions. Searches of standard citation indexes turn up little or no rigorous empirical work. Two main factors may explain this neglect. First, shame is not generally regarded by emotion researchers following Panksepp (1998) as being among the basic emotions expressed in mammals generally. This comports with a tendency among emotion researchers to presume that shame requires cognitively complex reflexive representation; but a valid basis for this presumption is elusive, since there is no reason why an animal might not experience the emotion based on cues in conspecific behavior. Second, shame in humans is reasonably assumed to be a proximate indicator of fear of loss of social status, so in modeling may be assimilated under this more general kind of cost. However, this side-steps the question of whether the emotional aversiveness of shame implies *additional* costs to normative violations, over and above those associated with expected status losses themselves. When we wonder whether human shame makes people uniquely vulnerable to low-cost social punishment, this is the question of importance.

As argued in Ross (2007), and less directly in Dunbar (1996/1998), possession of language, rather than a specially evolved emotional response disposition, may be the key human distinction here. A nonhuman animal's violation of a social norm is likely to be known about and remembered only by actual witnesses. By contrast, news of a person's transgressions can be spread widely through linguistic gossip and stored in the cultural memory of the whole community, thereby multiplying costs to the transgressor. Language also facilitates normative institutions of collective forgiveness: A person's violations can be almost

costlessly punished, but then the punishment can also almost costlessly be erased if word is passed around that gestures or actions of restitution have been made. This is socially efficient for collective punishers in an indirect way that amplifies the low *direct* cost of the punishment action to its administrators. Suppose that only high-cost punishment, such as physical violence, were available. In a society based on specialization of labor, the resulting incapacitation of the transgressor is a collective cost to the community's productivity. A similar point applies to ostracism. Forgiveness conventions that can be encoded linguistically allow effects of punishment on a norm violator's productivity to be cancelled after punishment has been observed to be effective, so the violator's contribution to the social product can potentially be fully restored.

The capacity for language likely *did* involve a special evolutionary adaptation in the neural structures of early hominids (Deacon 1997). But this may have been sufficient for the establishment of substantially enhanced effectiveness of low-cost punishment to defend communities against free-riding, if emotional discomfort at detecting signs of reduced social status were already in place. Given evidence to date, it seems that postulating any additional special adaptation is gratuitous. Evidently, however, this speculation resting mainly on the point that absence of evidence isn't evidence of absence should be investigated by searching experimentally for emotional distress in response to conspecific-observed norm violations in intelligent non-human social animals.

Strong reciprocity is not uncommon in the "wild"

doi:10.1017/S0140525X11001324

W. G. Runciman

Sociology, Trinity College, Cambridge CB2 1TQ, United Kingdom.
wgr@wgrunciman.u-net.com
http://www.trin.cam.ac.uk/index.php7pageid=538

Abstract: Guala is right to draw attention to the difficulty of extrapolating from the experimental evidence for weak or strong reciprocity to what is observed in the "wild." However, there may be more strong reciprocity in real-world communities than he allows for, as strikingly illustrated in the example of the Mafia.

Guala rightly emphasizes the difficulties involved in extrapolating from the experimental evidence for reciprocity (whether strong or weak) to the analysis of cooperation and punishment in the "wild." The experiments, which, as Guala says, are remarkably robust, have significantly increased our understanding of the psychology of trust and retaliation; and their extension to a wider range of subjects in different cultural and social environments has shown how far these may modify or redirect presumptively universal predispositions and preferences. But they do not by themselves generate testable predictions about the costs borne by strong reciprocators in real-world communities, which enhance the probability of the community's survival and reproduction.

Calculation of the costs of punishment in risk or effort is a very different exercise from calculating monetary gains and losses. Among the Bergdama, a decision that an offender is to be thrashed, expelled, or put to death "is reached casually round the camp fire, and if necessary the young men are then told to enforce it" (Schapera 1956, p. 87). There is presumably a cost to the young men so instructed, particularly if the offender may resist. But the young men may welcome the opportunity for the legitimate exercise of violence and anticipate a reward in enhanced prestige. This is not revenge of the kind that is more often triggered in small face-to-face societies by sexual jealousies and antagonisms. But nor is it weak reciprocity. Nor are the young men bounty-hunters rewarded in cash for the risks they incur laying hands on a fugitive. The punishment is (we must assume) effective in enforcing cooperation. But it is not easy to see what the experimental literature contributes to the understanding of what exactly is going on.

If Guala is right, it is plausible to suppose that punishment of defectors, free-riders, and cheats is cheapest either in small egalitarian societies where the sanctions are cultural (i.e., the information affecting behaviour in the phenotype is transmitted by imitation or learning) and weak reciprocity does most of the work, or in fully evolved states where the sanctions are social (i.e., the information is coded in rule-governed practices which define institutional roles) and the punishers are rewarded for punishing. There is very little net cost in punishment by ridicule and ostracism of offenders where behaviour is easily monitored, and very little net cost to punishers in a police state where they are paid both to detect and restrain offenders and to punish fellow-citizens who refuse to inform on them. But what about the intermediate cases?

In the institutions for collective action which Elinor Ostrom has studied to such good effect, Guala argues that the problem of non-cooperation is resolved by removing the obstacles in the way of non-costly punishment and that costs are incurred more in setting up these institutions than in punishing defectors. But "covenants without the sword" (Ostrom et al. 1992) are not covenants without sanctions. In the inshore fishery at Alanya in Turkey, the fishers are assigned to their locations by lot at the beginning of each season, and the list of locations is deposited with the mayor and gendarme. Cheating is difficult because it is easy for the other fishers to observe it and they have a common interest in ensuring that their own rights will not be usurped. But they have also to be "willing to defend their rights using physical means if necessary" (Ostrom 1990, p. 220). There is an element of weak reciprocity in the fishers' relationships with one another, and Ostrom reports that disputes are generally handled at the local coffeehouse. There also appears to be a fit with the model presented by Boyd et al. (2010), in which the total cost of punishing a free-rider declines as the number of punishers increases. But costs are still costs, and strong reciprocity is waiting in the wings, so to speak, to ensure that the covenant is renewed.

A striking example of strong reciprocity at work is the Mafia, as documented by Gambetta (2009). The population under study is as far as it is possible to imagine from a community of Good Samaritans. It consists of adult males unconstrained in their behaviour either by the cultural sanction of conventional moral disapproval or the social sanction of control by agents of the state. Its members depend on sustained and predictable cooperation among themselves. But this involves trusting people of the very kind from whom there is least reason to expect that trust will be forthcoming. The survival and reproduction of the organization is therefore critically dependent on the punishment of non-cooperators against whom the sanction is severe physical injury or assassination. Trust is sustained by a costly signalling system which evolves by variation and selection in the classic Darwinian manner to ensure that threats of punishment of defectors are credible and, where implemented, efficient. Non-punishers are required to demonstrate their trustworthiness by punishing when ordered to do so, and they face the same severe sanctions if they refuse.

Group selection then comes into operation as the "families" in which cooperation is most successfully replicated drive those in which it is less successfully replicated towards extinction. The "families" are impermeable to immigrants carrying different traits, and are relatively stable in composition. Mafiosi sometimes allow their close kin to be recruited by other "families," but this does not compromise within-group behavioural homogeneity: the incomers are recruited only because they are known to be

trustworthy. Differences between groups are partly random, but occasionally arise through the infiltration of an undercover agent of the state. Some groups are in environments where they are under more pressure than others from the state. But those with a larger proportion of more trustworthy punishers have higher survival rates.

Lab support for strong reciprocity is weak: Punishing for reputation rather than cooperation

doi:10.1017/S0140525X11000884

Alex Shaw and Laurie Santos

Department of Psychology, Yale University, New Haven, CT 06511.

Alex.Shaw@yale.edu **Laurie.Santos@yale.edu**

https://sites.google.com/site/alexshawyale/

Abstract: Strong reciprocity is not the only account that can explain costly punishment in the lab; it can also be explained by reputation-based accounts. We discuss these two accounts and suggest what kinds of evidence would support the two different alternatives. We conclude that the current evidence favors a reputation-based account of costly punishment.

Guala reviews the anthropological literature on costly punishment and convincingly argues that little evidence supports the notion that costly punishment is responsible for maintaining cooperation in small-scale societies. We agree with Guala's argument, but feel that it does not go far enough. Indeed, we think Guala could expand this critique to findings from the laboratory as well. In particular, we argue that costly punishment observed in the lab may not support a model of strong reciprocity either.

As the target article nicely reviews, proponents of strong reciprocity often use examples of laboratory-based costly punishment as evidence that cooperation evolved through strong reciprocity. Unfortunately, strong reciprocity is not the only account that can explain costly punishment in these laboratory settings. Another view that could account for the laboratory evidence is reputation-based models, in which costly punishment is favored by virtue of reputational gains from punishing (Price 2008; Santos et al. 2011). Under this view, individuals punish in order to signal some non-observable underlying quality, such as an understanding of social norms (Fessler & Haley 2003).

Given that both strong reciprocity and reputation-based accounts predict punishment in laboratory economic games, how can we distinguish between these two alternatives empirically? One method is to explore the nuanced predictions that each specific model might make. The major claim of strong reciprocity is that individuals punish in order to *increase cooperation*, and this claim makes two behavioral predictions. First, people should be especially likely to punish non-cooperators relative to other norm violators. Punishment should thus be directed more often at non-cooperation than other immoral actions (e.g., infidelity, incest) that are unrelated to cooperation. If individuals punish those who violate other sorts of norms just as severely as those who violate cooperation norms, this would suggest that punishment may not have evolved to promote cooperation specifically. The second prediction of strong reciprocity accounts is that cooperation should be more influenced by punishment performed by human agents than other types of punishment. To date, many experiments have shown that people increase cooperation when punishment is allowed (Fehr & Gächter 2002), but no studies have shown that people respond to punishment more when it is performed by agents rather than any negative contingency (Thorndike 1927). To test this prediction, researchers would need to set up an experiment where the probability of having one's payments reduced for defection is the same, but the punishment would come from either a person or from a computer algorithm. Under strong reciprocity, the punishment by human agents should be more likely to increase cooperation than other types of punishment and this differential influence of punisher type should be more pronounced for punishment of non-cooperators than other violations.

Reputation-based accounts also make specific predictions that would not be expected under strong reciprocity accounts. Specifically, these accounts predict that people should be especially likely to punish if doing so can improve their reputation with others and should be sensitive to cues related to being observed. Much evidence in the laboratory has confirmed these predictions. There is evidence that people give more to punish defections when their decision will be known by other participants than when their decision to punish will remain anonymous (Kurzban et al. 2007; Piazza & Bering 2008b). Additionally, individuals appear to improve their reputations by appropriately punishing non-cooperators; punishers are seen as more trustworthy and deserving of respect and are actually rewarded monetarily (Barclay 2006; Nelissen 2008). These pieces of evidence favor a reputation-based account.

There is, however, one piece of evidence that at first glance might appear to go against reputation-based accounts: People do still punish when anonymous (Henrich & Fehr 2003). Nonetheless, there is at least one way to explain punishment in anonymous one-shot interactions – people may have mechanisms for punishing others to improve their reputation and this psychology may misfire in economic games causing anonymous participants to still punish at low rates (Price 2008). The target article dismisses this misfiring account for punishment because people still take costs to punish even when they self-report that they understand the one-shot nature of the interaction, give less in one shot dilemmas than they do in sequential situations, and continue to give even after repeated trials.

This is a point where we disagree with the target article, as we feel that a misfiring explanation can account for participants' behavior. To better understand this view, consider an analogous argument in a different domain, that of mating strategies. Teenage boys often take costs to buy pornography, even though they would surely self-report that they understand that they cannot reproduce with the attractive centerfold (Hagen & Hammerstein 2006). This mating "misfiring" phenomenon is analogous to the performance of participants who punish at cost even though they report understanding the relevant aspects of anonymity. In the same way that pornography tricks men's well-designed mating psychology by providing images of attractive women that could provide a great mating opportunity in the real world, punishment studies may trick people's well-designed reputation psychology by providing clear norm violations that could provide a great reputation building opportunity if they happened in the real world. In both cases, people can tell the difference between the artificial (one-shot interaction/pornography) and the real thing (repeated interactions/real women), yet they still respond to the artificial stimulus and do so even after repeated trials. In the case of mating "misfiring" we don't demand a new psychological explanation (e.g., *strong eroticism*; Tooby et al. 2009), so it isn't clear that we should in the punishment case either.

A true understanding of the mechanisms underlying punishment in laboratory studies of cooperation will involve moving away from a focus on whether individuals take costs to punish others and instead investigating what cues influence one's willingness to punish. This focus on specific predictions in the lab and the target article's focus on investigating things more naturalistically will together move the field toward a better understanding of how punitive sentiments function in people's minds and in the broader human society as a whole.

Altruistic punishment as an explanation of hunter-gatherer cooperation: How much has experimental economics achieved?

doi:10.1017/S0140525X11000902

Robert Sugden

School of Economics, University of East Anglia, Norwich NR4 7TJ, United Kingdom.

r.sugden@uea.ac.uk

https://www.uea.ac.uk/eco/People_old/Faculty/rsugden

Abstract: The discovery of the altruistic punishment mechanism as a replicable experimental result is a genuine achievement of behavioural economics. The hypothesis that cooperation in hunter-gatherer societies is sustained by altruistic punishment is a scientifically legitimate conjecture, but it must be tested against real-world observations. Guala's doubts about the evidential support for this hypothesis are well founded.

Guala appraises the hypothesis of strong reciprocity in the light of anthropological evidence from hunter-gatherer societies. Much of his discussion focuses on a particularly distinctive implication of this hypothesis, namely, the existence of *altruistic punishment.* An individual A engages in altruistic punishment when she incurs costs to punish some other individual B for an action by B which is contrary to social norms *but not specifically directed at* A.

The discovery of the mechanism of altruistic punishment is an achievement of behavioural and experimental economics. As Guala documents, there is now an influential literature arguing that altruistic punishment plays a fundamental role in stabilising cooperation in hunter-gatherer societies. If that claim were true, it would suggest the hypothesis that modern humans have hard-wired preferences for altruistic punishment; and were that hypothesis confirmed too, the methods of behavioural economics would have led to a major discovery in the domain of natural science. But is the claim justified?

Guala's review of the anthropological evidence suggests that it is not. I cannot claim expertise in anthropology, but I find the review convincing and consistent with my reading of a collection of research papers selected by leading advocates of strong reciprocity (Gintis et al. 2005; Sugden 2007). It seems that economic cooperation in hunter-gatherer societies is mainly among kin and between directly reciprocating partners. Individual acts of "punishment" (perhaps better described as revenge) are usually in response to harmful behaviour directed at the revenge-taker, as in many sexual conflicts, and tend to be discouraged by third parties who fear the socially destructive effects of cycles of revenge. If one person violates a norm without directing the harm at particular others, punishment tends to be a collective act, structured so that the cost to any individual is very low. It is well known that gossiping and ridicule are common punishments in hunter-gatherer societies. Guala is particularly convincing in suggesting that these practices allow coalitions of punishers to form without there being any obvious ringleader, and hence with minimal cost to individuals.

In the rest of this commentary I accept Guala's assessment of the anthropological evidence, and focus on the apparent conflict between this evidence and experimental observations of altruistic punishment. Is this conflict symptomatic of flaws in the methodology of experimental economics?

It is useful to consider how the hypothesis of altruistic punishment originated. From at least the 1960s, economists recognised that public goods are sometimes supplied through voluntary contributions. Some explanations of this fact postulated that donors were motivated by individual incentives (e.g., Olson 1965). Others postulated non-selfish motivations – usually altruism, but positive reciprocity was also proposed (Sugden 1984). With the development of experimental economics, the explanation of voluntary contributions to public goods became a prominent research programme. In the experimental design most commonly used in this programme, subjects interact anonymously in a game with real monetary payoffs. Each subject has an incentive to free-ride, but all subjects gain if all contribute. Face-to-face social pressures are screened out as far as possible, with the aim of isolating and investigating non-selfish motivations (or "social preferences").

This line of research has led to three main conclusions. First, a significant proportion of individuals make positive contributions, contrary to the assumption of self-interest. Second, individuals' contributions tend to be positively correlated with one another, as implied by positive reciprocity. Third, as the game is repeated, the rate of contribution decays. The best explanation of these findings seems to be that they result from interaction between two types of individuals in the population – free-riders and positive reciprocators. The positive reciprocators gradually withdraw from cooperation as they find insufficient reciprocation from others (e.g., Bardsley & Moffatt 2007).

Fehr and Gächter (2000a) initiated a new line of research by establishing that the tendency for contributions to decay can be overturned if, after each round of the public good game, each subject has the opportunity to impose costly punishments on individual others. For this mechanism to work, there must be some subjects with an "altruistic" preference for punishing free-riders, but Fehr and Gächter showed theoretically that there can be high and stable rates of contributions even if the proportion of such individuals is quite small and their preference for punishment is quite weak. There is now a large body of experimental evidence showing that, if the cost of punishing is low relative to the cost of being punished and if those being punished do not have the option to retaliate, high rates of contributions *are* sustained. The implication is that, *if placed in this experimental environment*, a sufficient proportion of individuals reveal a sufficiently strong preference for altruistic punishment for cooperation to be stabilised.

I agree with Guala that there has been real scientific progress here. The mechanism by which altruistic punishment can support cooperation is a genuine discovery which grew out of a programme of sound experimental research. But I agree too that there are many reasons for caution about extrapolating from these experiments to cooperation problems in real life. In particular, the anonymised experimental environment filters out mechanisms by which face-to-face contact might inhibit punishment; instead, it channels negative affective responses into punishment.

Understood as a brave conjecture inspired by experimental research, the hypothesis that hunter-gatherer cooperation relies on altruistic punishment is scientifically legitimate. But the hypothesis must be tested against real-world observations; and if it fails, it fails. Good science does not always succeed.

Punishing for your own good: The case of reputation-based cooperation

doi:10.1017/S0140525X11001336

Claudio Tennie

Max Planck Institute for Evolutionary Anthropology, 04013 Leipzig, Germany.

tennie@eva.mpg.de www.claudiotennie.de

Abstract: Contrary to Guala, I claim that several mechanisms can explain punishment in humans. Here I focus on reputation-based cooperation – and I explore how it can lead to punishment under situations that may or may not be perceived as being anonymous. Additionally, no particular mechanism stands out in predicting an excess of punishment under constrained lab conditions.

In explaining costly punishment in general, Guala's article focuses on (cultural) group selection, and subsequently "strong

reciprocity." But other mechanisms can also lead to punishment and which might only appear to be altruistic. These mechanisms include: kin selection; mutualism; pseudo-reciprocity, and direct reciprocity (see overview in Bshary & Bergmüller 2008). There is also general reputation-based cooperation, where "helpful and harmful acts [are returned] in kind" – even by third-parties (Nowak & Sigmund 2005), and/or where net-benefits accrue to both sender and receiver via the logic of costly signaling theory (see Barclay & Willer 2007; Nelissen 2008). In this commentary, I concentrate on such general reputation-based cooperation.

Within a reputation-based account, other-benefiting costs are reimbursed by an increase in that individual's reputation – which is beneficial in the long run. Imagine an agent acts as an altruistic punisher. In the long run, she may benefit since her punishing will increase her reputation, which in turn will render it more likely that she will benefit through the actions and choices of others.

In general, a good reputation can increase the likelihood of receiving future help (Wedekind & Braithwaite 2002) and of gaining effective partners (Sylwester & Roberts 2010) – while it decreases the likelihood of punishment ("bad" free-riders are punished) and shunning (Panchanathan & Boyd 2004). Push and pull.

Importantly, reputation might be established through several routes (Russell et al. 2008): by direct interaction, by eavesdropping (e.g., A observes B and C interact), and by reporting (gossip, e.g., A later reports to D). It is possible that reputation-based cooperation may be sensitive only to the amount of the cost paid (Nelissen 2008) but not to the form of the cost (e.g., helping vs. punishing; cf. Semmann et al. 2004). However, a reputation for being a punisher may also bring special (or additional) reputational benefits based on the threat developed through past punishment (dos Santos et al. 2011).

If the lab setup allows for reputation to be formed or passed on, then this (positively) influences general group beneficial behavior (Milinski et al. 2002; Wedekind & Milinski 2000). Reputation then also allows for individual benefits gained for punishment (Barclay 2006; Nelissen 2008); and punishing also increases with larger audiences (Kurzban et al. 2007). Thus, reputation is an important factor in such games. But how does reputation-based cooperation hold up in the light of punishing in anonymous lab situations? Following and extending others' work, I suggest that reputational concerns may still be key here.

Any perception of anonymity can be wrong: some audience may be observing after all (Frank 1988). Thus, either the perceived anonymity is real ("factual anonymity," i.e., no audience) or not ("false anonymity," i.e., some audience). Benefits in false anonymity are standard benefits in the reputation system: increased reputation (e.g., Barclay 2006). An actor should thus estimate the likelihood of the perceived anonymity being factually false, and should subsequently be more likely to punish the likelier it seems that some audience is present after all. As long as there is an overall net-gain across situations, this approach should work. But, of course, such an estimate will sometimes be wrong, and the actor may thus sometimes punish even in factual anonymity. Some have proposed that this explains punishment in anonymous lab settings, that is, that it may be ultimately due to a somewhat fixed, but outdated, adaptation of the past (and possibly helped by a general rarity of factual anonymous situations in the past). All this then falls under the label of the "Big Mistake hypothesis" (Richerson & Boyd 2005) – an evolutionary logic embraced by several authors (Burnham & Johnson 2005; Hilbe & Sigmund 2010; Tennie et al. 2010; Trivers 2004). In addition, subjects may still come to regard the test situation as a falsely anonymous one today (a "nagging suspicion" may remain; see Nowak & Sigmund 2005). In both cases, punishment in factual anonymity may still be due to intended self-interest in a reputation based system.

These explanations may still assume that any act of punishment in factual anonymity must always be associated with a net-cost for a reputation-based cooperator. But there might even be a net-benefit. As already mentioned, reputation can be passed on in the absence of audiences – through gossip. In this way even a punisher observed by no one may later benefit – namely, if the act of punishment enables him to later convince others that he indeed punished. This may be possible: his "bragging" will likely be more convincing if in fact he did punish, because (at least some) listeners can detect lies (Ekman & O'Sullivan 1991). Additionally, it is cognitively less costly to tell the truth (Paglieri 2007), and these costs may help outweigh the costs of the punishment. Finally (depending on the sort of punishment), others may observe the effects of his punishment and may therefore be able to independently validate his reports. Granted, in modern economic games these effects may be negligible, but the general Big Mistake logic might apply also here.

The scenario just described may explain the general presence of punishment in anonymous lab situations, but what about the apparently excessive frequency of punishment in the lab reported by Guala? Guala points out that punishment is often the main way in which participants can influence others' behavior in such economic games. The reason why we do not see equally high levels of punishment under field conditions may therefore simply be that real life offers more behavioral options that can influence others' behavior than the lab usually does. In addition, punishment seems to be a less preferred option in general (maybe even a "last resort") – Guala mentions the fear and cost of retaliation (see also Barclay 2006; Hill et al. 2009). In sum, special conditions in the lab (lack of choice; regardless of preference) in comparison to the field (availability and preference for other behavioral options) could explain the discrepant findings of punishment frequencies reported in Guala's article.

ACKNOWLEDGMENT

The author thanks Keith Jensen for helpful comments.

What we need is theory of human cooperation (and meta-analysis) to bridge the gap between the lab and the wild

doi:10.1017/S0140525X11000872

Paul A. M. Van Lange, Daniel P. Balliet and Hans IJzerman

Department of Social and Organizational Psychology, VU University, Amsterdam, 1081 BT Amsterdam, The Netherlands.

pam.van.lange@psy.vu.nl **dp.balliet@psy.vu.nl**
h.ijzerman@psy.vu.nl
http://www.paulvanlange.com
http://h.ijzerman.googlepages.com

Abstract: This commentary seeks to clarify the potential discrepancy between lab-based and field data in the use and effectiveness of punishment to promote cooperation by recommending theory that outlines key differences between the lab and field, such as the shadow of the future and degree of information availability. We also discuss a recent meta-analysis (Balliet et al. 2011) that does not support all conclusions outlined in Guala's target article.

As shown in experimental research in laboratories, introducing a system whereby people can punish non-cooperators at a cost to self effectively promotes cooperation (e.g., Fehr & Gächter 2002; for a recent meta-analysis, see Balliet et al. 2011). Such lab-based research findings are now being challenged in Guala's provocative target article, which reviews evidence from anthropology (and some other disciplines) to reach the conclusion that people may not engage in costly punishment to encourage cooperation outside the laboratory. Although Guala identifies an important discrepancy between lab-based and field data, we argue that this discrepancy, at least as it currently exists, is less problematic than the target article suggests. We

emphasize two solutions to understanding this discrepancy: namely, theory and meta-analysis.

From the lab to the "wild": Why we need theory. Social interactions during lab-based studies differ substantially from interactions in the "wild." According to interdependence theory (Kelley et al. 2003; Van Lange & Rusbult 2012), such differences in social interactions are likely rooted in profound differences in the situations that people face. What might be key differences in interdependence between situations in the lab and the wild? One important difference is the one between single-trial versus repeated interaction (i.e., *temporal structure*). In small communities, people typically know each other, have a history of social interaction experiences, and may have formed strong attachments to (some) other members of their community. In contrast, in the lab, researchers have almost exclusively relied on strangers. Moreover, there is almost always an interdependent future ahead of people in the wild, which is not typical for lab research.

Another situational feature of interdependence theory is *information availability*. There is little doubt that the "rules of the game" are clearer in the lab than in the wild. For example, in the lab, it is often specified how costly an action is for a person, but in the real life such costs are often less tangible or known. Psychologically, the lack of information about others' preferences, magnitude of costs, and the like, challenge important processes such as those linked to risk, uncertainty, and interpersonal trust.

The next question is, of course, this: In what ways might temporal structure and information availability be important to our understanding whether costly punishments are effective, or why they are used only sporadically in the field? Temporal structure is important because it is linked to the history of social interaction experiences as well as the anticipated future of social interaction. The history is important to relationship development, including growth or decline in feelings of attachment, and trust, along with its manifestations and embodiments (e.g., IJzerman & Koole 2011). The future is important because it may trigger the "shadow of the future" (Axelrod 1984); that is, an implicit or explicit mindset that tends to promote cooperation, so that punishment is less often called for (Van Lange et al. 2011). In light of the future, what might people do? It seems reasonable to expect that people first start to communicate, or to gossip, as Guala notes, before seeking punishment in any more material, or tangible sense.

Information availability may also affect punishments of non-cooperators. Costly punishment may be less likely to be used if one is not completely confident that another person intentionally acted as a non-cooperator. For example, in everyday life, unintended errors (called *noise*) are quite commonplace, in that external barriers prevent a person from translating his or her cooperative intentions into cooperative action (e.g., Van Lange et al. 2002). Given that people at some level are likely to realize that particular outcomes are not intentional, it is likely that people might give each other the benefit of the doubt. And in light of incomplete information, people may seek information about the intentionality of an action instead of immediately punishing a non-cooperator.

From the lab to the "wild": Why we need meta-analysis. While theory provides the concepts and logic, a meta-analysis provides the comprehensive databank – from the lab and the field – that should allow for rigorous conclusions. In this regard, it is interesting that Guala's narrative review of the lab studies reaches the conclusion that "'costly' punishment works only if it costs relatively little" (sect. 12, para. 3). However, in a recent meta-analysis of incentives and cooperation including 187 effect sizes from both psychology and economic studies, Balliet et al. (2011) found that costly punishments (and rewards) were more effective at encouraging cooperation, compared to when punishments were free to administer. They interpreted their findings in terms of interdependence theory (Kelley et al. 2003; Van Lange & Rusbult 2012), noting that people translate the costs of administering punishment as a strong indicator of benign intent. Hence, the conclusion about the cost of punishment reached by Guala may be inaccurate because of a restricted survey of the literature.

Concluding remarks. Interdependence theory provides a taxonomy of situations that can be fruitfully used to explicate the key differences in situational structure between the lab and the field. This theory provides much needed insight into the key elements of situations that affect social interaction processes in dyads and groups. We have suggested the relevance of temporal structure and information availability for understanding the discrepancy between lab and field research on costly punishment, although other features (e.g., asymmetries in dependence) might be crucial as well.

We hope that field data about the topic of punishment (and reward) and cooperation grow, and that their operationalizations match those of the lab. Likewise, we hope that researchers in the lab focus more strongly on key features of interdependence that characterize social dilemmas outside of the laboratory. Such efforts should make future meta-analyses even more informative. As Kurt Lewin noted, there is nothing more practical than a good theory (Lewin 1952, p. 169). Perhaps we may add: We need good data as well, as good data ultimately determine the contribution of a meta-analysis.

The social costs of punishment

doi:10.1017/S0140525X11001348

Pieter van den Berg, Lucas Molleman, and Franz J. Weissing

Theoretical Biology Group, Centre for Ecological and Evolutionary Studies, University of Groningen, 9700 CC Groningen, The Netherlands.

pieter.van.den.berg@rug.nl

l.s.molleman@rug.nl f.j.weissing@rug.nl

www.rug.nl/fmns-research/theobio

Abstract: Lab experiments on punishment are of limited relevance for understanding cooperative behavior in the real world. In real interactions, punishment is not cheap, but the costs of punishment are of a different nature than in experiments. They do not correspond to direct payments or payoff deductions, but they arise from the repercussions punishment has on social networks and future interactions.

We applaud Guala for pointing out that the results of punishment experiments cannot readily be generalized to "real-world" situations. However, we disagree with Guala's assertion that real-world punishment mechanisms such as ostracism and public ridicule are cheap or even costless. Instead, the costs of punishment can be very high, but they are of a different nature than their typical implementation in experiments suggests. In real-world interactions, the costs of punishment are usually not in terms of direct payoff deductions for the individuals carrying out the punishing. Instead, the effects of punishment on the punishers are more hidden and indirect, because they result from the repercussions of punishment behavior on social networks and social interactions. There are at least four reasons why such "social" costs of punishment can be substantial.

Punishment may have repercussions leading to a less favorable equilibrium with lower payoffs. In evolutionary ecology, one distinguishes between the direct and the ecological costs of a trait. For example, Strauss et al. (2002) discuss the costs of resistance to herbivory in plants. Direct costs of herbivory resistance can mostly be described in terms of resource allocation; resources allocated to defense mechanisms

cannot be allocated to growth or reproduction. The ecological costs of herbivory resistance are more long-term and indirect; examples are decreased attractiveness to pollinators or decreased competitive ability. The costs associated with punishment mechanisms such as ostracism may be distinguishable in a similar way. We agree with Guala that the direct short-term costs associated with ostracizing free-riders will often be low. However, on the longer term, there can be strong negative implications. Ostracized individuals may become desperados, causing a lot of trouble. They may resort to antisocial or criminal behavior, affecting the feeling of safety in their former group and necessitating protection measures. In the worst case, trust and cooperation break down. This way, the presence of ostracized individuals in the environment can lead to a new equilibrium with lower payoff levels than in the original state. Although there have been some experiments that include ostracism as an option (e.g., Maier-Rigaud et al. 2009; Masclet 2003), they do not accommodate those "ecological" costs.

Punishment in one type of interaction may have implications for different types of interaction. Economic experiments typically focus on a single type of interaction, such as a public goods game. If punishment is incorporated in these experiments, it can only affect behavior in that specific context. This is not in line with how behavior is structured in humans (and other animals). There is ample evidence that behavioral tendencies in one type of interaction are closely correlated with the behavior in quite different contexts. As shown by evolutionary models from the biological literature (e.g., Wolf et al. 2007; 2008), such correlation structures (called "behavioral syndromes" or "personalities") can be adaptive, even if the behavior in a particular type of situation may appear maladapted. For example, the tendency to show antisocial behavior in a public goods context may – for good reasons – be correlated with the tendency to actively participate in group defense when the group is facing an external challenge. Ostracizing individuals because of their behavior in a public goods context may therefore have harmful effects later.

Punishment may destroy established hierarchies and role patterns and lead to social unrest. The participants of a typical economic experiment do not know each other well and interact anonymously. In real life, many interactions take place in small communities where individuals do know each other, and are well aware of their place in the group. Individuals differ in relevant aspects (like age, expertise, or authority), and relationships between individuals (like leadership and social rank) have been settled in the past. Such patterning of a group due to well-established relationships between its members is important, because it reduces conflict and facilitates division of labor. Punishing an individual by social exclusion can break down such group structures, leading to social unrest. The re-establishment of stable social relationships can take a long time, and some individuals may end up in a worse position than they had before. Guala himself refers repeatedly to the work of Ostrom (1990), who has shown that stable group membership is one of the key predicting features making institutions for collective actions viable.

Punishment may have asymmetric effects, thus leading to tension between group members. Interactions in economic experiments are usually random. In contrast, real-world interactions take place in interaction networks that are often highly structured. This can be important, because group members may differ considerably in the way they are connected to a punished individual. Individuals will differ not only in the degree they suffer from the free-riding behavior of a specific individual, but also in the implications that punishment of that individual may have for them. Ostracizing an individual may have a small effect on group member A, while it severely affects the social network of group member B. The costs and benefits of punishing a particular free-rider can therefore be highly asymmetrical, leading to contrasting preferences between group members and, as a consequence, to social tension within the group.

If punishment were as cheap as Guala suggests, one would expect that individuals would readily punish defectors. In contrast, daily-life experience tells us that individuals are reluctant to punish free-riding group members. Denouncing others is often considered a bad habit, even if these others exhibit antisocial behavior. Groups of students assigned to a joint project, for example, are typically not only reluctant to punish free-riders, but even to call in an authoritative person (such as a professor) to resolve the situation. In fact, whistle-blowing is considered more a vice than a virtue, as young children are already being told by their parents or at school. This reluctance to apply seemingly cheap punishment is an indication that the hidden, long-term costs of punishment may be substantial. Economic experiments focusing exclusively on the direct costs of punishment are valuable, but they do not tell much about how cooperation is stabilized in human societies. For a complete understanding, the social costs of punishment should be taken seriously.

When the strong punish: Why net costs of punishment are often negligible

doi:10.1017/S0140525X11001427

Christopher R. von Rueden and Michael Gurven

Integrative Anthropological Sciences, Department of Anthropology, University of California, Santa Barbara, Santa Barbara, CA 93106-3210.

vonrueden@umail.ucsb.edu gurven@anth.ucsb.edu
http://sites.google.com/site/chrisvonrueden/home
http://www.anth.ucsb.edu/faculty/gurven/

Abstract: In small-scale societies, punishment of adults is infrequent and employed when the anticipated cost-to-benefit ratio is low, such as when punishment is collectively justified and administered. In addition, benefits may exceed costs when punishers have relatively greater physical and social capital and gain more from cooperation. We provide examples from the Tsimane horticulturalists of Bolivia to support our claims.

We agree with Guala that regulation of cooperation by punishment is infrequent and often low-cost, at least in small-scale societies. Analytical models and experimental studies suggest that solutions to cooperative dilemmas do not depend on direct punishment if individuals can opt out of unproductive partnerships (Aktipis 2004; Hauert et al. 2007) or assort with preferred cooperative partners whether kin (Hamilton 1964) or non-kin (Barclay & Willer 2007; Noe & Hammerstein 1994). Guala cites Wiessner (2005), who observes that !Kung who shirk their responsibilities are ignored more often than they are verbally punished. Among traditional Tsimane horticulturalists of Bolivia, most conflicts are between close kin and regular cooperative partners (von Rueden et al. 2009), who generally prefer reconciliation to revenge. Furthermore, defection among parties with few long-term shared interests is more often met with withdrawal and "voting with one's feet" than with punishment.

Guala does not distinguish second- from third-party punishment, but strong reciprocity theorists argue that both contribute to the maintenance of social norms (Fehr & Fischbacher 2004). There is no consensus over whether third parties often punish or punish "enough." In experimental games, third-party punishment is least common in small-scale societies (Marlowe et al. 2008), and third parties may be especially wary of becoming involved in serious conflicts. A Tsimane man committed

murder on two occasions, but punishment (public whipping) was administered only after the second murder. The community that sentenced and whipped him was not his resident community but a more acculturated community with more influential men. Nonpartisan members of the murderer's own community would not risk the threat of his retaliation.

Punishment occurs when there is minimal risk of (1) losing a valued exchange partner, (2) suffering reputational damage, or (3) provoking retaliation. For example, a low-status Tsimane man was in long-standing disputes with his neighbor over land and with his son-in-law over investment in his daughter. With few allies to support him, the man moved to another village with his family, with plans to return in a few months. The next day, the neighbor harvested the yucca from the man's field, and the son-in-law burned the man's house. The neighbor and son-in-law did not expect reputational damage or retaliation because they had strong kin support within the community, they could not be unambiguously identified as the punishers, and the punished man had few allies.

Guala identifies gossip as a low-cost alternative to direct punishment. Gossip can spread reputation-damaging information while obscuring the source of that information. Individuals may also gossip to gauge and build community support for punishment that is coordinated and more direct. As Guala argues, punishment that is coordinated carries less risk of retaliation and can be more effective at stabilizing collective action than distributed, individual acts of punishment (Boyd et al. 2010; Casari & Luini 2009). Among the !Kung, Wiessner (2005) found that most harsh criticism was delivered by a coalition, and coalition-based punishment was twice as likely to provoke conformity in the accused. Among the Tsimane, most conflicts are confined to the parties directly involved, but on occasion a small, informal gathering of men will act as third-party adjudicators. The most serious conflicts among the Tsimane, such as those involving physical violence, are sometimes discussed in community-wide meetings in more acculturated villages, where influential individuals will try to generate consensus concerning the relative guilt of the parties in conflict. The community may decide to inflict punishment, usually verbal censure, community service, or public whippings on rare occasions. One village has a de facto rule that the whipper not yet be a father; he has no risk then of his children being targets of vengeance.

Coordinated sanctioning, however, may not be necessary to explain why individuals punish free-riders and non-punishers. Another explanation, which Guala does not discuss, relies on inter-individual differences in formidability, endowments, or in the expected gains from successful cooperation. Individuals with greater physical or social capital can punish with less risk of retaliation and with greater efficacy, and those who anticipate greater relative gains from cooperation are more willing to absorb costs of punishment to achieve those gains. In general, inter-individual differences can be powerful catalysts of cooperation, transforming prisoner's dilemmas into mutualisms and resolving second-order dilemmas of who punishes (Olson 1965; Ruttan 2008). Among the Tsimane, 66% of adjudicated conflicts were arbitrated by men in the top 10% of coalitional support within their community (von Rueden et al. 2009). These individuals can steer conflict outcomes in their favor and their actions are less likely to be challenged.

Inter-individual differences in the costs to punishing contribute to the establishment of leaders and followers. Collective action, particularly in large groups, often depends on leaders bearing the costs of coordination and punishment in return for a greater share of the spoils (Hooper et al. 2010). Tsimane men do not gain more direct material benefits from organizing collective fishing events or acting as leaders in face-to-face collective action games (von Rueden et al. 2010), but long-term reputational benefits may be non-trivial. Positive reputations can serve as insurance against times of need (Boone & Kessler 1999; Gurven et al. 2000) or as signals to mates and allies of quality or cooperative intent (Bliege Bird & Smith 2005). Where joint production is subject to greater economies of scale, such as in agricultural societies, coordination and punishment by leaders may pay even greater dividends.

We encourage more study of the role of inter-individual differences in the generation of punishment and cooperation. In the lab, players often feel equally entitled and motivated, while subject endowments are too often windfalls; these conditions rarely hold in natural settings. As Guala recognizes, context matters in shaping how social preferences impact behavior (Gurven & Winking 2008; List 2006; Wiessner 2009), so caution is required when making inferences from particular experimental games. Some experimental games, however, have introduced asymmetries into the effectiveness with which players punish (Nikiforakis et al. 2011), into decision-making authority over the distribution of public good shares (van der Heijden et al. 2009), or into initial player endowments, as a function of individual inputs to joint production (Konigstein 2000). With greater understanding of the pervasiveness of inter-individual differences and other cost-reducing conditions, punishment may not appear so altruistic after all.

Perspectives from ethnography on weak and strong reciprocity

doi:10.1017/S0140525X11000860

Polly Wiessner

Department of Anthropology, University of Utah, Salt Lake City, UT 84108.
wiessner@soft-link.com
http://www.anthro.utah.edu/faculty/wiessner.html

Abstract: To add ethnographic perspective to Guala's arguments, I suggest reasons why experimental and ethnographic evidence do not concur and highlight some difficulties in measuring whether positive and negative reciprocity are indeed costly. I suggest that institutions to reduce the costs of maintaining cooperation are not limited to complex societies.

Guala's target article makes a most welcome contribution to the discussion of strong reciprocity, crossing disciplines to compare the findings of economic experiments and ethnographic evidence from small-scale egalitarian societies, "in the wild." It comes as no surprise to anthropologists that the two do not concur; cooperation in the wild is tamed by emotions accompanying kinship, a factor lost in experiments that hinge on anonymity (Wiessner 2009). Moreover, the one-shot material consequence of punishment in experiments in no way parallels the multi-shot social consequences of the same in real life. Grudges from punishment, particularly by third parties, are infinitely retrievable and accrue; punishment begs retribution, petty or pernicious, that so disrupts cooperation.

Significant also is Guala's point that positive and negative strong reciprocity are not the flip side of the coin. Cooperation in small-scale societies is driven largely by benefits, not by blows, whether social or physical. Strong punishers are not rewarded for their sacrifices while strong positive reciprocators are revered. Among the Kalahari Bushmen, pushing back to regulate weak reciprocity provides the spice of daily life, but frequent harsh punishers, particularly the few third party punishers, are despised and called *tshi n!ai* or "biting thing." In 62% of Bushman conversations that involved some social sanctioning where the camp leader was present, the leader refrained from sanctioning in order to save his clout for subsequent mediation (Wiessner 2005).

Whether positive and negative reciprocity are costly and thus truly "strong" is difficult to measure in the field. For the six out of 124 cases of sanctioning among the !Kung Bushmen that I

evaluated as costly, three were over land rights, two over sexual promiscuity, and one over deposing an aging leader. The costs of reacting immediately were most likely to have been lower than the consequences of no action for the land and sex disputes. More than five coalition members shared costs of sanctioning in all cases because the offender threatened the entire community. In contrast, acts that appear to be positive strong reciprocity are frequent in small-scale societies. I say "appear" because positive strong reciprocity builds reputation or symbolic capital that may be cashed in years later (Bourdieu 1977). A few studies suggest that the costs of generous acts are balanced out by benefits in the long-run (Smith et al. 2003; Wiessner 2002).

Sophisticated institutions to manage sustained cooperation are not unique to complex societies. For example, compensation, a form of restorative justice, practiced in Melanesia (Lemonnier 1990; Strathern 1971; Trompf 1994; Wiessner & Tumu 1998) puts a neat twist on strong negative and positive reciprocity. The victim's kin require the offender and his kin to pay compensation for insult, injury, homicide, or destruction of property; else they threaten violent retribution. If the offending parties come up with a generous payment, they receive fame and acclaim, turning punishment to positive strong reciprocity. Lasting ties are renewed and new ones may be created to produce a win–win situation. People in small-scale societies are well aware of the costs of punishment; institutions to reduce those costs did not wait for the Leviathan.

Author's Response

Strong reciprocity is real, but there is no evidence that uncoordinated costly punishment sustains cooperation in the wild

doi:10.1017/S0140525X1100166X

Francesco Guala

Department of Economics, University of Milan, 20122 Milan, Italy.
francesco.guala@unimi.it
http://users.unimi.it/guala/index.htm

Abstract: I argue in my target article that field evidence does not support the costly punishment hypothesis. Some commentators object to my reading of the evidence, while others agree that evidence in favour of costly punishment is scant. Most importantly, no rigorous measurement of cost-benefit ratios in the field has been attempted so far. This lack of evidence does not rule out costly punishment as a cause of human cooperation, but it does pre-empt some overconfident claims made in the past. Other commentators have interpreted my article as an anti-experimental pamphlet or as a flat denial of the existence of pro-social motives – which it was not intended to be. While we have enough data to establish the existence (and theoretical relevance) of strong reciprocity motives, I argue in this response that their efficacy (and policy relevance) has not been demonstrated.

R1. Introduction

Strong reciprocity theory is controversial, so it is not surprising that the target article generated a diverse set of commentaries. This diversity suggests that we are still a long way from resolving the main disagreements, but it also confirms that any attempt to clarify the empirical status of the theory should be welcome at this stage of the debate. I am grateful to all the commentators for their feedback: I agree with a number of points they have made – and when there is agreement, I will not dwell upon it; but even when I disagree, the commentaries will give me the opportunity to clarify the main theses of my article and to try to articulate what remains to be done.

Although the target article is mainly about the empirical status of the costly punishment story, it contains implicitly an alternative account of human cooperation. This view is in some respects unconventional, which may have caused some confusion. It is worth stating succinctly before I engage with the commentaries in detail. I believe that the following four claims can be true simultaneously, and that they explain the available evidence better than alternative explanations:

1. Strong negative reciprocity is a real proximate cause of human behaviour, and may *indirectly* promote cooperation.
2. Punitive motives in particular sustain institutions that administer sanctions, satisfying the moralistic preferences of the cooperative members of society.
3. To be viable, however, these institutions typically reduce the costs of sanctions and in particular prevent the eruption of feuds triggered by uncoordinated punishment.
4. Uncoordinated costly punishment thus is unlikely to be a *direct* mechanism sustaining cooperation in the wild. Successful societies either find ways to administer coordinated punishment at low cost, or abstain from punishing free-riders, as documented by anthropologists.

This view is unconventional in that it admits the existence of strong reciprocity, while downsizing its explanatory and practical relevance. It also draws a subtle distinction between the reality of pro-social motives and the channels through which they may (or may not) promote sociality in the wild. It is not contradictory to recognize the reality of a phenomenon and yet acknowledge that it is of limited explanatory importance. Fundamentalists from both sides – who argue either that strong reciprocity does not exist, or that it exists and is an important cause of cooperation – will be disappointed, but this is what the evidence suggests right now, as I have argued in my target article and will continue to argue in this response. I begin my replies responding to a couple of commentators who share my distaste for fundamentalism.

R2. Is strong reciprocity a straw man?

The first question to ask is whether I have got my polemical target right. Since my argument depends on the existence of different research programmes in the study of human cooperation, misunderstanding what these programmes are about would inevitably invalidate my project from the start. **Henrich & Chudek** believe that I have constructed a straw man – "an empty set of 'strong reciprocity theorists'" – that does not reflect a real scientific divide in the study of human sociality. They claim that while weak reciprocity is a class of theoretical models, strong reciprocity refers to a set of empirical regularities for which a number of theoretical explanations have been proposed, many of which are actually of the weak reciprocity kind. So in comparing

the weak and strong programmes, I would be making a category mistake – comparing apples with oranges, so to speak.

In my target article I define strong reciprocity as the propensity to reply "nice" to nice actions and "nasty" to nasty actions (sect. 1, para. 1), even if this entails a net cost for the individual agent. The strong reciprocity *programme*, as I understand it, aims at explaining various aspects of human sociality using models that incorporate strong positive or strong negative reciprocity motives (sometimes called "social preferences") among their premises. The *costly punishment hypothesis*, which is the main focus of my target article, is an important element of strong reciprocity theory, that is, the idea that uncoordinated costly sanctions supported by strong negative reciprocity efficiently discipline free-riders and protect positive reciprocators in social dilemma games.

Any label, of course, is bound to be imprecise: In science, originality is prized, and there are few incentives to repeat exactly the story told by other scholars. Cooperation studies, moreover, are highly interdisciplinary. Because scientists working in different fields have different agendas, we should expect a certain amount of heterogeneity within any programme. Like others before me, I have therefore taken the "strong reciprocity" label only as a useful ideal type. However, the fact that a number of scholars are willing to defend the claims that I have attributed to strong reciprocity theorists (see, e.g., the commentaries of **Gächter**, **Gintis & Fehr**, and **Bowles, Boyd, Mathew, & Richerson** [**Bowles et al.**]) demonstrates that I have not built a straw man. Many of these claims concern the best way to *explain theoretically* the regularities observed in punishment experiments. So **Henrich & Chudek**'s interpretation of strong reciprocity as primarily an empirical enterprise, uncommitted to a distinctive theoretical/explanatory strategy, is not shared by other scientists in the strong reciprocity camp.

This does not mean, of course, that **Henrich & Chudek**'s preferred interpretation is illegitimate. As researchers who have made important contributions to the study of human sociality, they are entitled to endorse an interpretation that differs from that of "core" strong reciprocity theorists. But they should not deny that such a core exists and that it has become influential over the last decade. Quite simply: Henrich & Chudek are not "purists" (and wisely so, in my view).

R3. Are weak and strong reciprocity mutually exclusive?

As evidence that I mischaracterize their programme, **Henrich & Chudek** cite several papers in which strong reciprocity theorists appeal to weak reciprocity mechanisms (such as, kinship, or reputation) to explain human cooperation. But nowhere in my target article do I say that weak and strong reciprocity are incompatible, nor do I attribute this claim to strong reciprocity theorists. The controversial issue is whether strong reciprocity has a significant effect on cooperation, over and beyond the effect of weak reciprocity.

There is plenty of evidence in support of weak reciprocity, as several commentators highlight. Cultural factors (e.g., norms) are favoured by many (e.g., **Boehm**, **Henrich & Chudek**, **Read**, and **von Rueden & Gurven**) and I fully agree that this is where future research should concentrate. I also agree with **Feinberg, Cheng, & Willer** (**Feinberg et al.**), **Ross**, **Shaw & Santos**, **Tennie**, and **von Rueden & Gurven** that reputation and gossip are crucial to enforce norm-compliance. **Baumard** and **Wiessner** point out that compensation of the victim is used in many societies to eliminate free-riding advantages and to recoup the costs of punishment. Von Rueden & Gurven notice that the costs of punishment vary considerably across individuals, due to individual differences such as wealth or physical strength. Finally, **Casari**, **Dreber & Rand**, **Ferguson & Corr**, and **Ostrom** highlight that a large amount of cooperation does not require punishment at all. I endorse all these comments: Weak reciprocity has many resources to explain spontaneous cooperation in small societies and cooperation regulated by common-pool institutions.

Still, **Henrich & Chudek** are right to say that no evidence in favour of weak reciprocity, by itself, counts as evidence against strong reciprocity. My primary goal in the target article is not to argue that weak reciprocity is the *only* mechanism sustaining cooperation in human societies. It is, more modestly, to point out that the data that are routinely cited by supporters of strong reciprocity do not provide genuine support to the costly punishment story. These data can be equally well explained by weak reciprocity models, and it is a basic principle of confirmation theory that a body of evidence supports hypothesis A over B only if the evidence is more likely to be observed under the assumption that A (rather than B) is true.

It is important to step back and consider objectively the situation that my article was trying to address: Anyone who has read the strong reciprocity literature of the past few years has derived the clear impression that (1) costly punishment can solve dilemmas of cooperation in the lab, and (2) there is a substantial amount of field evidence in favour of the costly punishment hypothesis. I took the latter claim for granted myself, until I began accidentally to review the evidence on my own. What I discovered did not fit with the claims made by strong reciprocity theorists, and this convinced me that we should ask for better evidence before we take the costly punishment story on board. My view is in line with the caution expressed by those scholars (like **Boehm** and **Ostrom**, for example) who have spent many years gathering and reviewing data on cooperation in the wild. Both recognize that the available field data can be explained differently, and they warn against overinterpreting the evidence to fit a preconceived theoretical framework.

R4. Can field data solve the riddle of cooperation?

An important methodological thesis of my article is that laboratory data alone cannot solve the riddle of cooperation. It is easy to take this as a general anti-experimental argument, which is perhaps why some commentators felt the need to stress that field data alone are not enough either. I agree: My article was never meant to be an anti-experimental treatise, as **Casari** and **Nikiforakis** seem to have interpreted. I am very fond of laboratory experiments, a methodology that has greatly enriched the toolbox of social science and human biology (Guala 2005). My suggestion is, again more modestly, that at

this particular juncture we will gain more by combining what we have learned in the laboratory with the message of field studies. This in turn will lead to more (and better) experiments, which, together with new field data, will drive progress in this difficult but fascinating field of research. Nikiforakis and Casari already use this eclectic approach in their work. **Bereby-Meyer** and **Pisor & Fessler** provide further examples of how it is possible to bring more realism in experiments, or how to incorporate in experimental designs some features that are typical of the real-world circumstances in which cooperation takes place. My trivial point is that in order to run more realistic experiments, we need to know more about cooperation in the wild.

Experimental research may be driven by theoretical questions, experimental questions, field questions, or (in various proportions) by all three. So far, there has been a tendency for the reciprocity debate to be overly concerned with the first two types of questions. My article was intended to promote a more balanced approach and to re-direct experimental research towards field questions. (**Ostrom**'s research is a good model in this respect.)

R5. Does ethnographic evidence support the costly punishment hypothesis?

Gintis & Fehr write that "anthropologists have confirmed that strong reciprocity is indeed routinely harnessed in the support of cooperation in small-scale societies." Without further argument or justification, this is just the same claim they have repeatedly made in previous publications, and which my target article challenges. Surprisingly, Gintis & Fehr cite in support the *same* ethnographic literature that I claimed they have misreported in previous work. The only new entry is Henrich et al. (2010a), which is not a field study but reports the results of cross-cultural experiments – perpetuating one of the misleading confusions between field and experimental data that I try to dispel in my article.

Of the old literature, **Gintis & Fehr** keep citing the work of Boehm (1984, 2000) and Wiessner (2005, 2009). In my article I argue that the evidence reported in these studies does not support the costly punishment story. (One of the articles by Wiessner [2009], by the way, says so explicitly.) The commentaries published in this issue of *BBS* support my interpretation: **Wiessner** agrees that "experimental and ethnographic evidence do not concur" (see her Abstract), and **Boehm** similarly claims that "the costs do not necessarily fit with assumptions made in models that consider punishment to be altruistic" (Abstract). Other anthropologists (e.g., **Baumard**, **von Rueden & Gueven**, and **Read**) argue that there are plausible alternative readings of the evidence. (Read in fact says that I do not go far enough in my discussion of the "disjunction between experimental and real conditions.")

Finally (and ironically) **Gintis & Fehr** refer to **Henrich & Chudek**'s commentary as a source of evidence in favour of the importance of costly punishment in small societies. But as we have seen (sect. R2), Henrich & Chudek subscribe to a much broader interpretation of the strong reciprocity programme, in which costly punishment does not play a prominent role. In fact, in their commentary Henrik & Chudek explicitly say that "models relying on DCP [diffuse costly punishment] ... are not consistent with how norms are actually stabilized in small-scale societies."

Bowles et al. pursue a better strategy, citing new evidence in favour of the costly punishment account. The study of Turkana warfare by Mathew and Boyd (2011) is interesting and confirms that sanctions can be important to enforce cooperation. The version of this paper that I have seen, however, does not include any analysis (quantitative or qualitative) of the *costs* of punishment. So the claim of Bowles et al. that "punishing takes time and effort and may damage valuable social relationships" seems unsupported by the paper they cite. On the contrary, Mathew and Boyd (2011) provide evidence in favour of the importance of coordination, coalitional punishment, and the imposition of fees on free-riders – all mechanisms that reduce the individual costs of punishment and the social dilemma problem. A similar story seems to apply to Meggitt's (1962) study of the Walbiri and Strehlow's (1970) study of Aranda foragers. As Bowles et al. explicitly say, in both cases the community plays a prominent role in the decision to sanction, appointing the punisher and protecting from retaliation. In the Aranda case, retaliation seems to have been occasionally carried out – which is consistent with a large body of anthropological literature. But the point here is not the existence of punishment or violence per se, which everyone agrees is all too common in small societies. The point is whether punishing is costly (because of the risk of retaliation) and at the same time is able to improve, rather than damage, social relations. I have not seen yet a set of quantitative data that answers this question in a convincing fashion, nor have scholars such as **Boehm** who have reviewed the ethnographic literature more widely. Overall, I doubt that we will find an old study that was designed just in such a way as to answer this question. What we need are especially customized new measurements, where all the obvious confounds have been estimated and tested using rigorous statistical techniques.

R6. Have I overlooked some field data in favour of strong reciprocity?

In an interdisciplinary debate of this kind it is very difficult, perhaps impossible, to review all the relevant literature. So I am not surprised that many commentators have identified holes in my survey. **Johnson**, for example, mentions a field experiment by Gerber et al. (2008) on voters' turnout, where the threat of naming (and, presumably, shaming) non-voters raised turnout by eight percent. The experiment clearly suggests that people care about their reputations, but, as far as I can see, it does not say anything about the cost of punishment and people's willingness to incur such costs for the sake of enforcing cooperation.

Casari says that costly punishment is still practised in Trentino, the region at the centre of his research on the *Carte di regola*. He mentions damages to young grapevines carried out at night, but not enough detail is provided to figure out what these stealth expeditions are really about. Are they meant to enforce cooperation in the pursuit of a common good? Or are they just petty

jealousies among neighbours? Are these *individual* initiatives (perhaps part of ongoing feuds) or *coordinated* actions backed up by the whole community? These questions are crucial, because as I have said, the existence of punishment is not at issue here, nor, similarly, in the ethnographic literature. The issue is whether such punishment is costly to individual punishers, and whether it sustains or disrupts social cooperation.

R7. Is evidence for strong reciprocity hard to find because costly punishment is rare?

A related methodological issue raised by various commentators concerns the intrinsic difficulty of observing costly punishment in action. **Gächter**, **Gintis & Fehr**, **Johnson**, and **Nikiforakis** point out that negative reciprocity mechanisms are most effective when they work as *deterrents*, that is, when it is not necessary to use them frequently. This is crucial because, as Balliet et al. (2011) show in a recent meta-analysis, there is a tension between two aspects of punishment devices: Costly sanctions are more effective at raising cooperation (they send a stronger message of disapproval, presumably); but they also tend to undermine efficiency if applied too often (see also the commentary by **Van Lange, Balliet, & IJzerman** [**Van Lange et al.**]). Low-frequency sanctions may be the only viable costly punishment regimes in the long run.

Before I address the argument in more detail, let me highlight that appealing to rarity amounts to a significant retreat with respect to previously published claims: Whereas in earlier writings strong reciprocity theorists reported the existence of costly punishment as an established fact, we are now told that it is an elusive phenomenon, and that we should not expect to see very much of it when we look at field data. This looks suspiciously like a "heads I win, tails you lose" kind of argument. Even though absence of evidence is not evidence of absence, it hardly counts as evidence of presence either.

Having said that, is the retreat empirically justified? **Gächter** discusses in some detail the results of an experiment showing that, in equilibrium, punishment is rare. His clarifications are particularly welcome, given that the published article (Gächter et al. 2008) is a one-page report that leaves much unstated. In the experiment, subjects play a Public Goods game with punishment for 50 consecutive rounds (an unusual length in experimental economics) *with the same partners*. Notice that this is not a particularly good setting for strong reciprocity, because reputation-building is likely to play some role. Gächter and colleagues find significantly more cooperation in a condition with punishment, than in a no-punishment condition. They also find higher net earnings overall, in contrast with previous (shorter) experiments where punishment did not pay.

There is, however, a significant drop during the very last period (a classic end-effect), where average earnings reach the same level as in the no-punishment condition. The drop is caused by two factors: a decrease of contributions, and an increase of punishment in the last round of the game. This suggests that the shadow of the future is important: The subjects who defect in the last round presumably do not expect to be punished because they believe (incorrectly) that the others will not consider punishment worthy.

The emphasis on error is quite important, as highlighted in **Dreber & Rand**'s commentary. What looks like an equilibrium when error is not permitted, may turn out to be unstable in a stochastic environment. Uncertainty is likely to play an important role in real-world environments – recall that one of the complaints of strong reciprocity theorists is that the almost perfect monitoring required by folk theorems is unrealistic. **Bereby-Meyer** notices that the introduction of uncertainty in Ultimatum games reduces the rate of rejections significantly (see also **Van Lange et al.** for related comments). This might explain why punishment is observed only rarely in the field, but it is a rather different type of explanation from **Gächter**'s: If people give others the benefit of the doubt, free-riding becomes more profitable and sanctions *less* effective. In section 13 of the target article I explain how successful institutions help solve this problem, by coordinating monitoring and resolving whatever uncertainties there may be (e.g., on the interpretation of rules). While the existence of such institutions is almost certainly backed up by strong reciprocity motives, their smooth functioning relies on weak reciprocity mechanisms that guarantee long-term profitability, sustainability, and efficacy.

R8. Does punishment have to be uncoordinated?

Some commentators criticize my assumption that strong reciprocity sanctions ought to be "diffuse" or uncoordinated. **Gintis & Fehr** and **Bowles et al.** criticize me explicitly for this, but the same point is implicit in **Henrich & Chudek**'s claim that punishment must be understood more broadly than I do in my target article. I confirm that I do make this assumption; but is it really unjustified? My "narrow" characterization is based on the empirical fact that in the overwhelming majority of experiments punishment is indeed uncoordinated. As I point out in the main article, this was not true of seminal studies such as Yamagishi (1986) or Ostrom et al. (1992), but for a long time this particular feature of their designs was not appreciated by strong reciprocity theorists. Now a new wave of theoretical and empirical work (e.g., Boyd et al. 2010; Casari & Luini 2009; Ertan et al. 2009) is reintroducing coordinated punishment in the debate – a positive development in my view. But it is important to realize that coordination in real-world institutions has the very important function of *reducing the cost* of punishment. Coordination brings two important benefits: It legitimizes the sanction, which is backed up by the (implicit or explicit) assent of the group's majority; and it also reduces the likelihood that the sanction will be counter-punished. These two mechanisms remedy an important defect of standard uncoordinated punishment, but go against the grain (and the spirit) of strong reciprocity theory, with its emphasis on self-regulation and altruism.

R9. A small cost is still a cost, but is there any evidence of it?

Bowles et al. say that it is illegitimate to suppose that the cost of punishment ought to be large. But how large is "large" in this context? The 1:3 ratio between cost and inflicted damage that is used in many experiments is

unrealistic for situations in which punishment can be retaliated by equally strong individuals. But even the 1:3 ratio generates inefficient outcomes (e.g., Egas & Riedl 2008). In such circumstances, either cooperation is bound to collapse, or people must devise cheaper ways to enforce it. I suspect that both cases are common, but the study of successful resilient institutions suggests that if there are superior alternatives to uncoordinated costly punishment, people tend to exploit them. So I agree with **Casari** that one important reason why costly punishment is not frequently observed in the field is that people find better ways to enforce cooperation.

Still, cheaper punishment is not necessarily costless punishment, and even small costs are inconsistent with weak reciprocity models. I agree, but no field study (especially those routinely cited by strong reciprocity theorists) includes a rigorous attempt to calculate the cost-benefit ratio of punitive behaviours in the wild. Let me stress again that I am not saying that there is evidence in favour of the zero-cost hypothesis. As anthropologists know all too well, it is very difficult to collect evidence on cost-benefit ratios outside the lab. A major problem is that while the costs may be immediately evident, the benefits (in terms of enhanced reputation, access to sexual mates, etc.) are likely to be delayed and diffuse. That is why the literature on reciprocity abounds with anecdotal, non-quantitative examples.

But many anecdotal "costs" that are routinely cited during talks, seminars, conversations, and even printed articles, are not relevant for the reciprocity debate. "Psychological costs" (**Gächter** and **Adams & Mullen**), for example, are irrelevant unless they reflect some underlying material cost, because psychic distress does not cause a comparative disadvantage and therefore does not create a free-rider problem. One can speculate that psychological negative reactions (e.g., anger, moral disgust) were selected for some reason in the ancestral past, and therefore must reflect some evolutionary advantage. Because the relevant time-scale for the debate on human reciprocity is in my view the medium term of cultural evolution, I am reluctant to engage in these evolutionary speculations. And in any case the issue cannot be decided on such grounds: As the debate on evolutionary psychology has taught us, an emotional reaction that was selected under different pressures may systematically "misfire" and be a real cause of current behaviour even though it does not provide any current cost-benefit advantage (**Shaw & Santos**).

Van den Berg, Molleman, & Weissing [**van den Berg et al.**] cite costs generated by ostracism that can easily be overlooked, like the creation of predatory outcasts ("desperados") or the disruption of social relations that are crucial for a well-functioning group. While I agree with them that further research is required on these costs and their quantitative impact, I should point out that their existence is well known to ethnographers. **Boehm**, for example, describes various mechanisms observed in small societies that have the effect of distributing the costs of sanctions over the members of the group and of alleviating some side-effects of punishment. Kinsmen are chosen to act as punishers or peacemakers; the identity of executioners is kept secret, or the group as a whole acts as killer. Boehm also notices that such mechanisms are determined culturally and situationally, which reduces the problem of (genetic) free-riding.

I do not have the expertise to comment on the importance of these cultural mechanisms, but I have no doubt that we ought to study them in more detail. One important message of my target article is that it is time to abandon anecdotal evidence and move on to quantitative analysis. The assassination of mobsters mentioned by **Runciman**, unless articulated in further detail, belongs to the realm of the anecdote. As Gambetta (1993) explains convincingly, trust and reputation (i.e., weak reciprocity) are crucial cogwheels in the functioning of the Sicilian Mafia. And the very strategy of costly signalling mentioned by Runciman can be explained using standard game-theoretic models based on Nash equilibrium, in which the costs are recouped later in the game. Runciman is right, I believe, to say that in every successful institution "strong reciprocity is waiting in the wings … to ensure that the covenant is renewed." My purpose is not to deny that strong reciprocity motives exist (see also sect. R13 further on), but to point out that there is no evidence that they sustain cooperation by way of uncoordinated costly punishment in the wild.

R10. Are costs recouped via group selection?

A more radical strategy is to deny that the cost-benefit balance is important. **Henrich & Chudek** are the only commentators following this argumentative route, which is consistent with their ecumenical interpretation of strong reciprocity (see sect. R2). They argue that costs may be paid for via intergroup competition: An individual belonging to a highly cooperative group may be relatively disadvantaged with respect to a free-rider *within* her group, but this disadvantage may be recouped at a higher level if her group gains material (e.g., territorial) advantages through warfare.

This argument relies on group selection, which is itself a controversial theory in evolutionary biology. According to one interpretation, group selection models are just special cases of standard models based on inclusive fitness and kin selection, that apply when certain parameters take extreme values (e.g., when within-group competition is very low – see West et al. 2011, for a recent statement). Under this interpretation, then, **Henrich & Chudek** are right that weak and strong reciprocity explanations do not differ radically. However, all the objections to a costly punishment account of field data that I present in my article identify some mechanism (like reputation, coalitional punishment, etc.) that reduces the relative costs *within* the group. If the objections are sound, the free-rider problem may be negligible or non-existent, and there may be no need to recoup the costs at a higher level via group selection. This does not mean that competition between groups is not important, of course; only, that it might not solve *this* particular problem. (In fact, the opposite is likely to be true: Group selection works more smoothly if the free-rider problem within each group has already been solved using non-costly punishment mechanisms; see, e.g., Sober & Wilson 1998).

In my target article I do not put much emphasis on group selection because it plays an ancillary role in this debate. Scholars in both camps agree that at *some* level the costs have to be recouped. The contentious issue is where: If punishment is costly, then group selection has

a lot of work to do; if it is not, group selection may have an easier job or (perhaps) no job at all.

R11. Are experiments good predictors of field behaviour?

To support the external validity of experimental data, various commentators mention correlations between behaviour observed in laboratory settings (e.g., altruistic punishment) and related behaviour in real-life situations (e.g., participation to common projects, or consumption of common pool resources). Such correlations were beginning to be published when I was writing my target article, and therefore they did not receive the attention they deserved (cf. Henrich et al. 2010c; Rustagi et al. 2010). The strength and robustness of the correlations are crucial to warrant the use of experiments as measurement devices ("social thermometers," cf. Guala 2008). Moreover, this issue is strictly related with larger, controversial issues such as the relative importance of personality traits as opposed to situational factors in determining behaviour.

Bowles et al., **Ferguson & Corr**, **Henrich & Chudek**, and **Johnson** highlight the positive correlations as proof that laboratory behaviour predicts (at least partially) behaviour in the field. **Civai & Langus**, **Pisor & Fessler**, **von Rueden & Gurven**, and **Wiessner** in contrast highlight the *lack* of correlation found in other studies as evidence of the importance of contextual factors. My view is that we need a systematic analysis of when, where, and why such correlations obtain, before we can say anything general about the power of experiments as predictors of non-laboratory behaviour. Focusing on successes (e.g., positive correlations) may be justified at an early stage of research, when one is looking for surprising results, but at a later stage it must be supplanted by a quantitative assessment of successes and failures.

One plausible conjecture is that the external validity of experimental measures is highly dependent on how the experimental setting is interpreted by the participants. This is true of all experiments, regardless of the pool of subjects. **Civai & Langus** and **Güney & Newell** remind us that the results of Ultimatum and Dictator games vary with relatively small manipulations of the design. Adding real effort or "property rights" over the resource to be divided, for example, influences offers and rejections significantly. Egalitarianism is just one of several norms that can be triggered experimentally, and whose application depends on context. If a society recognizes that individual effort is to be rewarded, the effect of that norm can be observed experimentally by suitably modifying the design. The moral is that the design of experiments must fit what one intends to measure.

In the case of non-Western societies it is often hard to say what one is measuring. **Wiessner**, for example, notices that the anonymity precept of experimental economics creates a highly unusual environment for the members of small societies. While the very structure of Ultimatum or Public Goods games triggers familiar cues in Western subjects who are used to bargaining and cooperation, it is difficult to imagine what goes on in the minds of people whose economic activities do not depend on trust and negotiations with strangers. Notice that the argument here is not that these games do not trigger any real-world norm of behaviour (every game situation has to be interpreted, after all), but that they may cue heterogeneous behaviours that are highly dependent on contextual factors. This would explain why cross-cultural experiments have generated more varied results, compared with those performed with Western subjects (see, e.g., Henrich et al. 2010c). This is a key point especially for the interpretation of the ethnography of cooperation, and my position is that claims based on experimental correlations should be treated with extreme care until we know more about them.

Having said that, let me emphasize that I never meant to claim that the results of experimental games have no external validity. On the contrary, I believe they do in a number of cases. In fact, it would be surprising to find no correlation between the behaviour in and out of the lab. My external validity worry is different: Uncoordinated costly punishment may be a bad solution to the problem of cooperation because in realistic environments it creates more problems than it solves. That is why societies have found alternative ways to sustain cooperation and to harness the natural impulse to sanction free-riding. The problem is not that strong negative reciprocity occurs in the lab only: On the contrary, because it is a real force everywhere, it has to be carefully managed, channelled, and if necessary suppressed.

R12. Should we talk about ultimate causes only?

Several commentators have highlighted problems with the way in which evolutionary, economic, and psychological explanations are mingled in the reciprocity debate. **Dos Santos & Wedekind**, for example, accuse strong reciprocity theorists of confusing proximate and ultimate explanations, while **Barclay** points out that an advantageous (selfish) behaviour from an ultimate perspective need not be selfish from a psychological (proximate) perspective. Both commentaries claim that weak reciprocity theories are concerned with ultimate mechanisms only, and therefore cannot be criticized for their failure to account for the pro-social motives of cooperative agents.

I generally agree with the spirit of these comments. Philosophers of biology have introduced important distinctions between "psychological," "economic," and "biological" altruism that have helped clarify the debate, and which should always be kept in mind (Sober & Wilson 1998). The only point of (partial) disagreement is that the strong reciprocity programme in my view is not the main culprit regarding the mixing of proximate and ultimate explanations. The way I have characterized it, the weak reciprocity programme is *also* as a theory of proximate and ultimate causes. This is inevitable, once we decide to unify biological and economic approaches and to include standard game-theoretic accounts in the weak reciprocity camp. I understand that some biologists may be reluctant to make this move, but several social scientists and psychologists find it appealing.

This unification has rather unpleasant implications for weak reciprocity theory, though, because models based on selfish preferences and strategic reasoning are too limited to account for the variety of proximate causes of human behaviour (**Rosas**). One solution is to retreat to an "as if" interpretation of these models, and defend them as useful instruments based on unrealistic assumptions. Although there is an old instrumentalist tradition in economics, "as

if" interpretations have been used too often to shield theories from criticism. In contrast, models based on false principles should be modified to build better proximate models consistent with the spirit of weak reciprocity. The success of folk theorem–like explanations prompts us to ask how such idealized models can nevertheless be useful as stylized explanations – a question that has puzzled many scientists and philosophers since David Hume formulated it three centuries ago. But in search for better models one does not have to ditch the promising features of weak reciprocity explanations (like the emphasis on repeated play or reputation).

R13. Are pro-social motives real?

No reciprocity theorist today would claim that pro-social emotions (including anger at injustice, or punitive drives generally) are unreal. Similarly, no one would seriously argue that human behaviour is always calculative or strategic. Apart from psychopaths we are all (psychologically) pro-social, altruistic people. **Rosas** puts it nicely, saying that humans are psychologically unselfish, but biologically selfish creatures. **Civai & Langus**, **Jensen**, and **Ross** mention animal studies on emotions that may shed further light on the evolutionary origins of these mechanisms. Research in this area is just beginning to take off, to be sure, so it is not surprising that scholars disagree on the basic facts. (Reciprocity exists among animals, say Civai & Langus; but chimps do not display pro-social preferences in Ultimatum or Dictator games, according to Jensen). Following Ross, I suspect that until we have better data on animal emotions, this issue may be more usefully tackled by focusing on the mechanisms that amplify the negative consequences of bad reputation and, hence, explain the emergence of a distinctively human sensitivity to social emotions. *Language* has been for a long time the main suspect, so like Ross I believe that the key to solve the riddle of cooperation is culture.

Tennie adds that our cognitive limitations probably contribute to widen the domain of cooperative behaviour: Telling the truth, for example, is less costly than constantly strategizing. I agree whole-heartedly: The debate on reciprocity, as I see it, hinges on the interpretation and relative importance of subtle phenomena like these. An important issue is the *robustness* of pro-social emotions and behaviour to losses and repeated encounters. Another one is the flexibility of norms (like truth-telling, egalitarianism, etc.) to changes in strategic incentives. While friends of strong reciprocity see pro-social norms and emotions as very robust even outside the folk-theorem domain, weak reciprocity theorists are sceptical. The two approaches do not postulate radically different proximate causes, but disagree on their efficacy or robustness across various circumstances. The fact that the room for disagreement has been progressively reduced is testimony to the great work done by experimenters over the last decade, many of whom were inspired by strong reciprocity theory.

R14. Why does it matter?

As in my target article, I have left the most important issue for the very end. Cooperation studies are not just fascinating from a theoretical point of view but have potential policy implications as well. One reason why the interpretation of punishment experiments invites caution is that strong reciprocity models carry the risk of making cooperation appear too easy. I tend to read Hume's knavery principle in this light – as an antidote to complacency, rather than as expressing confidence in the correctness of the self-interest assumption.

Contemporary research on social capital highlights that individual pro-social tendencies ought to be nurtured and cannot be taken for granted. Putnam (2001), to cite a well-known study, shows that there is a strong link between continuous participation in the activities of the local community (weak reciprocity), on the one hand, and more general pro-social attitudes (e.g., altruism), on the other. The capacity to cultivate long-term relationships is correlated with people's willingness to cooperate outside the small circle of friends and family, and it is subject to medium-term cycles of growth and decay. All this suggests that the important levers for policy purposes lie *outside* the psychology of individuals, in the social structures that sustain and guide people's decisions in different circumstances. *Less individual psychology and more social science*, in a nutshell, would be my slogan for future research.

This invitation to caution is not meant to devalue strong reciprocity models or experiments. On the contrary, I believe that the strong reciprocity programme is important enough that we can look straight at its promises and its limitations. The question, "What mechanisms sustain cooperation (or can sustain cooperation) in some set of real-world conditions?" is in many respects separate and independent from theoretical questions concerning the existence of social preferences and the refutation of self-interest models. Success in one task does not imply success in the other (and "good science," **Sugden** reminds us, "does not always succeed").

Physicists have established the existence of different forces in nature (gravity, electromagnetism, weak and strong interactions, etc.). Nevertheless, they recognize that there is a wide gap between existence and explanatory power. There is no doubt that electromagnetism is real, or that it can be used to bring about astonishing effects in some conditions – heavy objects can be lifted in the air using electromagnetic forces, for example. But this does not mean that electromagnetism plays a significant role in making airplanes fly. To understand why airplanes fly, and to improve their performance, air pressure and fluid mechanics are much more important than electromagnetism. Something similar might be true of strong reciprocity. There is a wide gap between theoretical relevance and application, and we should better acknowledge that strong reciprocity theory has not bridged it yet.

References

[The letters "a" and "r" before author's initials stand for target article and response references, respectively]

Aktipis, C. A. (2004) Know when to walk away: Contingent movement and the evolution of cooperation. *Journal of Theoretical Biology* 231:249–60. [CRvR]

Aldrich, J. H. (1995) *Why parties?* University of Chicago Press. [TJ]

Alexander, R. D. (1974) The evolution of social behaviour. *Annual Review of Ecological Systematics* 5:325–83. [MdS]

Alvard, M. (2003) Kinship, lineage, and an evolutionary perspective on cooperative hunting groups in Indonesia. *Human Nature-an Interdisciplinary Biosocial Perspective* 14(2):129–63. [JH]

Anderson, C. & Putterman, L. (2006) Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. *Games and Economic Behavior* 54:1–24. [NN]

André, J. B. & Baumard, N. (2011) The evolution of fairness in a biological market. *Evolution* 65:1447–56. [NB]

Andreoni, J. (1990) Impure altruism and donations to public goods: A theory of warm glow giving. *The Economic Journal* 100(401):464–477. [EF]

Andreoni, J., Erard, B. & Feinstein, J. (1998) Tax compliance. *Journal of Economic Literature* 36:818–60. Available at: http://www.jstor.org/stable/2565123. [aFG]

Archer, J. (2004) Sex differences in aggression in real-world settings: A meta-analytic review. *Review of General Psychology* 8:291–322. [EF]

Archer, J. & Benson, D. (2008) Physical aggression as a function of perceived fighting ability and provocation: An experimental investigation. *Aggressive Behavior* 34:9–24. [EF]

Arno, A. (1980) Fijian gossip as adjudication: A communication model of informal social control. *Journal of Anthropological Research* 36:343–60. [MF]

Atwater, L. E., Waldman, D. A., Carey, J. A. & Cartier, P. (2001) Recipient and observer reactions to discipline: Are managers experiencing wishful thinking? *Journal of Organizational Behavior* 22(3):249–70. [GSA]

Axelrod, R. (1984) *The evolution of cooperation.* Basic Books. [YB-M, aFG, PAMVL]

Axelrod, R. & Dion, D. (1988) The further evolution of cooperation. *Science* 242(4884):1385–90. [YB-M]

Axelrod, R. & Hamilton, W. D. (1981) The evolution of cooperation. *Science* 211:1390–96. Available at : http://www.sciencemag.org/cgi/content/abstract/sci;211/4489/1390. [aFG]

Balafoutas, L. & Nikiforakis, N. (2011) Norm enforcement in the city: A natural field experiment. *Mimeo.* [NN]

Balikci, A. (1970) *The Netsilik Eskimo.* Doubleday. [DRe]

Balliet, D., Moulder, L. B. & Van Lange, P. A. M. (2011) Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin* 137(4):594–615. Available at: http://dx.doi.org/10.1037/a0023489. [rFG, PAMVL]

Barclay, P. (2006) Reputational benefits for altruistic punishment. *Evolution and Human Behavior* 27:325–44. [AS, CT]

Barclay, P. (2010) *Reputation and the evolution of generosity.* Nova Science Publishers. [PB]

Barclay, P. (2011) The evolution of charitable behaviour and the power of reputation. In: *Applied evolutionary psychology*, ed. C. Roberts, pp. 149–72. Oxford University Press. [PB]

Barclay, P. & Willer, R. (2007) Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B: Biological Sciences* 274:749–53. [CT, CRvR]

Bardsley, N., Cubitt, R., Loomes, G., Moffatt, P., Starmer, C. & Sugden, R. (2009) *Experimental economics: Rethinking the rules.* Princeton University Press. [aFG]

Bardsley, N. & Moffatt, P. (2007) The experimetrics of public goods: Inferring motivations from contributions. *Theory and Decision* 62:161–93. [RS]

Baron, J., Gowda, R. & Kunreuther, H. (1993) Attitudes toward managing hazardous waste: What should be cleaned up and who should pay for it? *Risk Analysis* 13(2):183–92. [NB]

Baron, J. & Ritov, I. (2008) The role of probability of detection in judgments of punishment. Unpublished manuscript. [NB]

Batson, C. D. (1991) *The altruism question: Toward a social-psychological answer.* Erlbaum. [HG, KJ]

Baumard, N. (2010a) *Comment nous sommes devenus moraux: Une histoire naturelle du bien et du mal.* Odile Jacob. [NB]

Baumard, N. (2010b) Has punishment played a role in the evolution of cooperation? A critical review. *Mind and Society* 9:171–92. Available at: http://www.springerlink.com/content/16734k611l07p502/. [aFG]

Baumard, N. (2011) Punishment is not a group adaptation: Humans punish to restore fairness rather than to support group cooperation. *Mind and Society* 10(1):1–26. [NB]

Beersma, B. & van Kleef, G. A. (in press) How the grapevine keeps you in line: Gossip increases contributions to the group. *Social Psychological and Personality Science.* [MF]

Bendor, J. (1993) Uncertainty and the evolution of cooperation. *Journal of Conflict Resolution* 37(4):709. [YB-M]

Bendor, J. & Swistak, P. (1995) Types of evolutionary stability and the problem of cooperation. *Proceedings of the National Academy of Sciences USA* 92:3596–600. Available at: http://www.pnas.org/content/92/8/3596. [aFG]

Bendor, J. & Swistak, P. (1997) The evolutionary stability of cooperation. *American Political Science Review* 91:290–307. Available at: http://www.jstor.org/stable/2952357. [aFG]

Bereby-Meyer, Y. & Roth, A. E. (2006) The speed of learning in noisy games: Partial reinforcement and the sustainability of cooperation. *American Economic Review* 96(4):1029–42. [YB-M]

Berg, J., Dickhaut, J. & McCabe, K. (1995) Trust, reciprocity, and social history. *Games and Economic Behavior* 10:122–42. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0899825685710275. [aFG]

Bergstrom, T. C. (2002) Evolution of social behavior: Individual and group selection. *Journal of Economic Perspectives* 16:67–88. Available at: http://www.jstor.org/stable/2696497. [aFG]

Besnier, N. (1989) Information withholding as a manipulative and collusive strategy in Nukulaelae gossip. *Language and Society* 18:315–41. [MF]

Bicchieri, C. (2006) *The grammar of society: The nature and dynamics of social norms.* Cambridge University Press. [CC]

Bicchieri, C. & Zhang, J. (2010) An embarrassment of riches: Modeling social preferences in Ultimatum Game. In: *Handbook of the philosophy of economics*, ed. U. Maki, pp. 1–19. Elsevier. [CC]

Binmore, K. (1998) *Game theory and the social contract II: Just playing.* MIT Press. [aFG]

Binmore, K. (1999) Why experiment in economics? *The Economic Journal* 109(453):16–24. Available at: http://onlinelibrary.wiley.com/doi/10.1111/1468–0297.00399/abstract. [aFG]

Binmore, K. (2005) *Natural justice.* Oxford University Press. [aFG]

Binmore, K. (2006) Why do people cooperate? *Politics, Philosophy and Economics* 5:81–96. Available at: http://ppe.sagepub.com/content/5/1/81. [aFG]

Bliege Bird, R. & Smith, E. A. (2005) Signaling theory, strategic interaction, and symbolic capital. *Current Anthropology* 46(2):221–48. [CRvR]

Blount, S. (1995) When social outcomes aren't fair: The effect of causal attributions on preferences. *Organizational Behavior and Human Decision Processes* 63:131–44. [CC]

Bochet, O., Page, T. & Putterman, L. (2006) Communication and punishment in voluntary contribution experiments. *Journal of Economic Behavior & Organization* 60(1):11–26. [SGä]

Boehm, C. (1984) *Blood revenge: The enactment and management of conflict in Montenegro and other tribal societies.* University of Pennsylvania Press. [CB, HG, rFG]

Boehm, C. (1999) *Hierarchy in the forest: The evolution of egalitarian behavior.* Harvard University Press. [HG, aFG, ACP]

Boehm, C. (2000) Conflict and the evolution of social control. *Journal of Consciousness Studies* 7:79–183. (Special issue on *Evolutionary origins of morality*, ed. L. Katz). [CB, rFG]

Boehm, C. (2008) Purposive social selection and the evolution of human altruism. *Cross-Cultural Research* 42:319–52. [CB]

Boehm, C. (2011) Retaliatory violence in human prehistory. *British Journal of Criminology* 51:518–34. [CB]

Boehm, C. (in press) *Moral origins: The evolution of altruism, virtue, and shame.* Basic Books. [CB]

Boone, J. L. & Kessler, K. (1999) More status or more children: Social status, fertility reduction, and long-term fitness. *Evolution and Human Behavior* 20:257–77. [CRvR]

Borges, B. F. J. & Knetsch, J. L. (1997) Valuation of gains and losses, fairness, and negotiation outcomes. *International Journal of Social Economics* 24:265–81. [aFG]

Bourdieu, P. (1977) *Outline of a theory of practice.* Cambridge Studies in Social Anthropology. Cambridge University Press. [PW]

Bowles, S. & Gintis, H. (2002) Behavioural science: Homo reciprocans. *Nature* 415:125–28. Available at: http://www.nature.com/nature/journal/v415/n6868/full/415125a.html. [aFG]

Bowles, S. & Gintis, H. (2004) The evolution of strong reciprocity: Cooperation in heterogeneous populations. *Theoretical Population Biology* 65(1):17–28. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0040580903001163. [MdS, AD, HG, aFG]

Bowles, S. & Gintis, H. (2005) Prosocial emotions. In: *The economy as an evolving complex system III*, ed. E. B. Lawrence & N. D. Steven, pp. 339–66. Santa Fe Institute. [HG]

Bowles, S. & Gintis, H. (2011) *A cooperative species: Human reciprocity and its evolution.* Princeton University Press. [SB, HG]

Boyd, R., Gintis, H. & Bowles, S. (2010) Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science* 328(5978):617–20. Available at: http://www.sciencemag.org/cgi/content/full/328/5978/617. [SB, HG, arFG, JH, WGR, CRvR]

Boyd, R., Gintis, H., Bowles, S. & Richerson, P. (2003) The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences USA* 100(6):3531–35. Available at: http://www.pnas.org/content/100/6/3531. [NB, SB, AD, SGä, aFG, JH,]

Boyd, R. & Richerson, P. (1990) Group selection among alternative evolutionarily stable strategies. *Journal of Theoretical Biology* 145:331–42. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0022519305801134, [aFG]

Boyd, R. & Richerson, P. (1992) Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology* 13(3):171–95. Available at: http://linkinghub.elsevier.com/retrieve/pii/016230959290032Y. [aFG, JH]

Briggs, J. L. (1970) *Never in anger: Portrait of an Eskimo family.* Harvard University Press. [aFG]

Bromley, D. W. (1992) *Making the commons work: Theory, practice, and policy.* Institute for Contemporary Studies. [MC]

Brosnan, S. F. & de Waal, F. B. M. (2003) Monkeys reject unequal pay. *Nature* 143:297–99. [CC]

Bshary, R. & Bergmüller, R. (2008) Distinguishing four fundamental approaches to the evolution of helping. *Journal of Evolutionary Biology* 21:405–20. [CT]

Burlando, R. M. & Guala, F. (2005) Heterogeneous agents in public goods experiments. *Experimental Economics* 8:35–54. Available at: http://www.springerlink.com/content/u38gh84100pk3235/. [aFG]

Burnham, T. C. & Johnson, D. D. P. (2005) The biological and evolutionary logic of human cooperation. *Analyse and Kritik* 27:113–35. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.169.3915&rep=rep1&type=pdf [aFG, CT]

Buss, D. M. & Duntley, J. D. (2008) Adaptations for exploitation. *Group Dynamics* 12:53–62. [EF]

Cabral, L., Ozbay, E. & Schotter, A. (2011) Intrinsic and instrumental reciprocity: An experimental study. *Mimeo.* [NN]

Camera, G. & Casari, M. (2009) Cooperation among strangers under the shadow of the future. *American Economic Review* 99(3):979–1005. [MC]

Camerer, C. (2003) *Behavioral game theory: Experiments in strategic interaction.* Princeton University Press. [YB-M, aFG]

Camerer, C. F. & Fehr, E. (2004) Measuring social norms and preferences using experimental games: A guide for social scientists. In: *Foundations of human sociality*, ed. J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr & H. Gintis, pp. 55–95. Oxford University Press. [aFG]

Cardenas, J. C., Stranlund, J. & Willis, C. (2000) Local environmental control and institutional crowding-out. *World Development* 28:1719–33. Available at: http://dx.doi.org/10.1016/S0305–750X(00)00055–3. [aFG]

Carlsmith, K. M. & Darley, J. M. (2008) Psychological aspects of retributive justice. *Advances in Experimental Social Psychology* 40:193–236. [GSA]

Carlsmith, K. M, Darley, J. M. & Robinson, P. H. (2002) Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology* 83(2):284–99. [NB]

Carpenter, J. (2007) The demand for punishment. *Journal of Economic Behavior and Organization* 62(4):522–42. [NN]

Carpenter, J. & Seki, E. (2011) Do social preferences increase productivity? Field experimental evidence from fishermen in Toyama Bay. *Economic Inquiry* 49(2):612–30. [SGä]

Carpenter, J. P., Daniere, A. G. & Takahashi, L. M. (2004) Cooperation, trust, and social capital in Southeast Asian urban slums. *Journal of Economic Behavior and Organization* 55:533–51. Available at: http://dx.doi.org/10.1016/j.jebo.2003.11.007. [aFG]

Casari, M. (2007) Emergence of endogenous legal institutions: Property rights and community governance in the Italian Alps. *Journal of Economic History* 67(1):191–226. Available at: http://journals.cambridge.org/abstract_S0022050707000071. [MC, aFG]

Casari, M. & Luini, L. (2006) *Peer punishment in teams: Emotional or strategic choice?* Purdue University Economics Working Papers, No. 1188, Department of Economics, Purdue University. [MC]

Casari, M. & Luini, L. (2009) Cooperation under alternative punishment institutions: An experiment. *Journal of Economic Behavior and Organization* 71(2):273–82. Available at: http://dx.doi.org/10.1016/j.jebo.2009.03.022. [MC, SGä, arFG, CRvR]

Casari, M. & Plott, C. R. (2003) Decentralized management of common property resources: Experiments with a centuries-old institution. *Journal of Economic Behavior and Organization* 51(2):217–47. Available at: http://dx.doi.org/10.1016/S0167–2681(02)00098–7 [MC, aFG]

Chagnon, N. A. (1968/1992) *Yanomamö: The fierce people,* 6th edition. Holt, Rinehart & Winston. (Original work published in 1968). [aFG]

Chagnon, N. A. (1988) Life histories, blood revenge, and warfare in a tribal population. *Science* 239(843):985–92. Available at : http://www.sciencemag.org/cgi/content/abstract/sci;239/4843/985. [aFG]

Chaudhuri, A. (2011) Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature. *Experimental Economics* 14(1):47–83. [SGä]

Cheng, J. T., Fournier, M. A. & Di Domenico, S. I. (2007) Status and affiliation: The psychological rewards that drive gossip behavior. Poster presented at the Society for Interpersonal Theory and Research's 10th Annual Convention, Madison, Wisconsin, June 19, 2007. [MF]

Cherry, T. L., Frykblom, P. & Shrogren, J. F. (2002) Hardnose the dictator. *American Economic Review* 92:1218–21. [ŞGü]

Chhatre, A. & Agrawal, A. (2009) Tradeoffs and synergies between carbon storage and livelihood benefits from forest commons. *Proceedings of the National Academy of Sciences USA* 106(42):17667–70. [EO]

Chudek, M. & Henrich, J. (2010) Culture-gene coevolution, norm-psychology, and the emergence of human prosociality. *Trends in Cognitive Sciences* 15(5):218–26. [JH]

Ciarocco, N. J., Sommer, K. L. & Baumeister, R. F. (2001) Ostracism and ego depletion: The strains of silence. *Personality and Social Psychology Bulletin* 27:1156–63. [GSA]

Cinyabuguma, M., Page, T. & Putterman, L. (2005) Cooperation under the threat of expulsion in a public goods experiment. *Journal of Public Economics* 89:1421–35. [HG, aFG]

Cinyabuguma, M., Page, T. & Putterman, L. (2006) Can second-order punishment deter perverse punishment? *Experimental Economics* 9(3):265–79. Available at: http://dx.doi.org/10.1007/s10683-006-9127-z. [AD]

Civai, C., Corradi-Dell'Acqua, C., Gamer, M. & Rumiati, R. I. (2010a) Are irrational reactions to unfairness truly emotionally-driven? Dissociated behavioral and emotional responses in the Ultimatum Game task. *Cognition* 114:89–95. [CC]

Civai, C., Corradi-Dell'Acqua, C., Rumiati, R. I. & Fink, G. R. (2010b) Disentangling between self- and fairness-related mechanisms in the Ultimatum Game: An fMRI study. Paper presented at the Second Meeting of the Federation of European Societies of Neuropsychology (ESN), September 22–24, 2010, Amsterdam, The Netherlands. [CC]

Clutton-Brock, T. & Parker, G. (1995) Punishment in animal societies. *Nature* 373(6511):209–16. [NB]

Coleman, E. (2009) Institutional factors affecting ecological outcomes in forest management. *Journal of Policy Analysis and Management* 28(1):122–46. [EO]

Coleman, E. & Steed, B. (2009) Monitoring and sanctioning in the commons: An application to forestry. *Ecological Economics* 68(7):2106–13. [EO]

Cooper, D. J. & Dutcher, E. G. (2009) The dynamics of responder behavior in ultimatum games: A meta-study. Working Paper, Florida State University. Available at: https://mywebdav.fsu.edu/djcooper/research/dynresponder.pdf. [aFG]

Croson, R. T. A. (1996) Information in ultimatum games: An experimental study. *Journal of Economic Behavior & Organization* 30(2):197–212. [YB-M]

Cubitt, R., Drouvelis, M. & Gächter, S. (2011) Framing and free riding: Emotional responses and punishment in social dilemma games. *Experimental Economics* 14(2):254–272. [SGä]

Dal Bó, P. (2005) Cooperation under the shadow of the future: Experimental evidence from infinitely repeated games. *American Economic Review* 95:1591–604. [AD]

Dal Bó, P. & Fréchette, G. R. (2011) The evolution of cooperation in infinitely repeated games: Experimental evidence. *American Economic Review* 101(1):411–29. [AD]

Dana, J., Weber, A. W. & Kuang, J. X. (2007) Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory* 33:67–80. [CC]

Darley, J. M., Carlsmith, K. M. & Robinson, P. (2000) Incapacitation and just deserts as motives for punishment. *Law and Human Behavior* 24(6):659–83. [NB]

Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R. & Smirnov, O. (2007) Egalitarian motives in humans. *Nature* 448:794–96. [ACP]

Dawkins, R. (1976) *The selfish gene.* Oxford University Press. [aFG]

Dawkins, R. (1976/2006) *The selfish gene* (30th Anniversary edition). Oxford Paperbacks. (Original work published in 1976). [PB]

Deacon, T. (1997) *The symbolic species.* Norton. [DRo]

Denant-Boemont, L., Masclet, D. & Noussair, C. (2007) Punishment, counter-punishment and sanction enforcement in a social dilemma experiment. *Economic Theory* 33(1):145–67. Available at: http://www.springerlink.com/index/J7356X79300876J7. [AD, aFG]

de Quervain, D. J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A. & Fehr, E. (2004) The neural basis of altruistic punishment. *Science* 305:1254–58. Available at: http://www.sciencemag.org/cgi/content/full/305/5688/1254. [aFG]

de Waal, F. B. M. (1996) *Good natured.* Harvard University Press. [DRo]

de Waal, F. B. M., Luttrell, L. M. & Canfield, M. E. (1993) Preliminary data on voluntary food sharing in brown capuchin monkeys. *American Journal of Primatology* 29:73–78. [KJ]

dos Santos, M., Rankin, D. J. & Wedekind, C. (2011) The evolution of punishment through reputation. *Proceedings of the Royal Society B: Biological Sciences* 278:371–77. [MdS, CT]

Dreber, A., Rand, D. G., Fudenberg, D. & Nowak, M. A. (2008) Winners don't punish. *Nature* 452(7185):348–51. Available at: http://www.nature.com/nature/journal/v452/n7185/abs/nature06723.html. [AD, MdS, aFG]

Dubreuil, B. (2010) Punitive emotions and norm violations. *Philosophical Explorations* 13:35–50. Available at: http://dx.doi.org/10.1080/13869790903486776. [aFG]

Dufour, C., Pele, M., Neumann, M., Thierry, B. & Call, J. (2009) Calculated reciprocity after all: Computation behind token transfers in orang-utans. *Biology Letters* 5:172–75. [CC]

Dufwenberg, M. & Kirchsteiger, G. (2004) A theory of sequential reciprocity *Games and Economic Behavior* 47:268–98. [CC]

Dunbar, R. I. M. (1996/1998) *Grooming, gossip and the evolution of language.* Harvard University Press. (Original publication date, 1996). [MF, aFG, DRo]

Dunbar, R. I. M. (2004) Gossip in evolutionary perspective. *Review of General Psychology* 8:100–10. [MF]

Eckel, C. C., Fatas, E. & Wilson, R. (2010) Cooperation and status in organizations. *Journal of Public Economic Theory* 12:737–62. [EF]

Egas, M. & Riedl, A. (2008) The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B: Biological Sciences* 275(1637):871–78. Available at: http://rspb.royalsocietypublishing.org/content/275/1637/871. [arFG, NN, ACP]

Ekman, P. & O'Sullivan, M. (1991) Who can catch a liar? *American Psychologist* 46:913–20. [CT]

Emery, N., Clayton, N. & Frith, C., eds. (2007) *Social intelligence: From brain to culture.* Oxford University Press. [DRo]

Ertan, A., Page, T. & Putterman, L. (2009) Who to punish? Individual decisions and majority rules in mitigating the free rider problem. *European Economic Review* 53:495–511. Available at: http://dx.doi.org/10.1016/j.euroecorev.2008.09.007. [SB, rFG]

Evans-Pritchard, E. E. (1940/1969) *The Nuer, a description of the modes of livelihood and political institutions of a Nilotic people.* Clarendon Press/Oxford University Press. (Original work published in 1940; 2nd edition, 1969, Oxford University Press). [NB]

Falk, A., Fehr, E. & Fischbacher, U. (2003) On the nature of fair behavior. *Economic Inquiry* 41:20–26. [KJ]

Falk, A., Fehr, E. & Fischbacher, U. (2005) Driving forces behind informal sanctions. *Econometrica* 73:2017–30. Available at: http://dx.doi.org/10.1111/j.1468-0262.2005.00644.x. [aFG]

Falk, A. & Fischbacher, U. (2005) Modeling strong reciprocity. In: *Moral sentiments and material interests*, ed. H. Gintis, R. Boyd, S. Bowles & E. Fehr, pp. 193–214. MIT Press. [aFG]

Falk, A. & Fischbacher, U. (2006) A theory of reciprocity *Games and Economic Behavior* 54:293–315. [CC, KJ]

Fehr, E. (2004) Human behaviour: Don't lose your reputation. *Nature* 432(7016):449–50. [JH]

Fehr, E. & Fischbacher, U. (2002) Why social preferences matter – The impact of non-selfish motives on competition, cooperation and incentives. *Economic Journal* 112:C1–C33. Available at: http://www3.interscience.wiley.com/journal/118940088. [aFG]

Fehr, E. & Fischbacher, U. (2003) The nature of human altruism. *Nature* 425:785–91. [MdS]

Fehr, E. & Fischbacher, U. (2004) Third-party punishment and social norms. *Evolution and Human Behavior* 25(2):63–87. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1090513804000054. [MdS, SGä, aFG, CRvR]

Fehr, E. & Fischbacher, U. (2005) The economics of strong reciprocity. In: *Moral sentiments and material interests*, ed. Gintis, R. Boyd, S. Bowles & E. Fehr, pp. 151–91 MIT Press. [aFG]

Fehr, E., Fischbacher, U. & Gächter, S. (2002) Strong reciprocity, human cooperation and the enforcement of social norms. *Human Nature* 13:1–25. [ACP]

Fehr, E. & Gächter, S. (2000a) Cooperation and punishment in public goods experiments. *American Economic Review* 90(4):980–94. Available at: http://www.jstor.org/stable/117319. [MdS, SGä, HG, aFG, TJ, AR, RS]

Fehr, E. & Gächter, S. (2000b) Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives* 14(3):159–81. [YB-M, HG]

Fehr, E. & Gächter, S. (2002) Altruistic punishment in humans. *Nature* 415(6868):137–40. Available at: http://www.nature.com/nature/journal/v415/n6868/abs/415137. [NB, MdS, SGä, HG, aFG, TJ, NN, ACP, AS, PAMVL]

Fehr, E., Gächter, S. & Kirchsteiger, G. (1997) Reciprocity as a contract enforcement device: Experimental evidence. *Econometrica* 65(4):833–60. [HG]

Fehr, E. & Gintis, H. (2007) Human motivation and social cooperation: Experimental and analytical foundations. *Annual Review of Sociology* 33:43–64. [YB-M, HG]

Fehr, E. & Henrich, J. (2003) Is strong reciprocity a maladaption? In: *Genetic and Cultural evolution of cooperation*, ed. P. Hammerstein, pp. 55–82. MIT Press. [JH]

Fehr, E., Kirchsteiger, G. & Riedl, A. (1993) Does fairness prevent market clearing? An experimental investigation. *Quarterly Journal of Economics* 108:437–59. Available at: http://www.jstor.org/stable/2118338 [aFG]

Fehr, E. & Leibbrandt, A. (2011) A field study of cooperativeness and impatience in the Tragedy of Commons. *Journal of Public Economics* 95(9–10):1144–55. [SB]

Fehr, E. & Rockenbach, B. (2003) Detrimental effects of sanctions on human altruism. *Nature* 422:137–40. [MdS]

Fehr, E. & Schmidt, K. M. (1999) A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* 114:817–68. [CC]

Fehr, E. & Schmidt, K. M. (2006) The economics of fairness, reciprocity and altruism – Experimental evidence and new theories. In: *Handbook of the economics of giving, reciprocity and altruism, vol. 1*, ed. S. Kolm & J. M. Ythier, pp. 615–91. Elsevier. [aFG]

Feinberg, M., Willer, R., Stellar, J. S. & Keltner, D. (2011) The existence and dynamics of prosocial gossip. Unpublished manuscript, University of California, Berkeley. [MF]

Ferguson, E., Atsma, F., de Kort, W. & Veldhuizen, I. (in press) Exploring the pattern of blood donor beliefs in first time, novice and experienced donors: Differentiating reluctant altruism, pure altruism, impure altruism and warm-glow. (Published online before print, August 16, 2011. DOI:10.1111/j.1537-2995.2011.03279.x). [EF]

Ferguson, E. & Chandler, S. (2005) A stage model of blood donor behaviour: Assessing voluntary behaviour. *Journal of Health Psychology* 10:359–72. [EF]

Ferguson, E., Farrell, K. & Lawrence, C. (2008) Blood donation is an act of benevolence rather than altruism. *Health Psychology* 27:327–36. [EF]

Ferguson, E., France, C. R, Abraham, C., Ditto, B. & Sheeran. P. (2007) Improving blood donor recruitment and retention: Integrating social and behavioral sciences agendas. *Transfusion* 47:1999–2010. [EF]

Ferguson, E., Heckman, J. J. & Corr, P. J. (2011) Personality and economics: Overview and proposed framework *Personality and Individual Differences* 51:201–209. [EF]

Fessler, D. M. T. & Haley, K. (2003) The strategy of affect: Emotions in human cooperation. In: *Genetic and cultural evolution of cooperation*, ed. P. Hammerstein, pp. 7–36. MIT Press. [AS]

Fischbacher, U., Gächter, S. & Fehr, E. (2001) Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters* 71:397–404. Available at: http://dx.doi.org/10.1016/S0165–1765(01)00394–9. [aFG]

Fong, C. (2001) Social preferences, self-interest, and the demand for redistribution. *Journal of Public Economics* 82:225–46. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0047272700001419. [aFG]

Francis, H. (1985) The law, oral tradition and the mining community. *Journal of Law and Society* 12:267–71. [MC]

Frank, R. H. (1988) *Passions within reason: The strategic role of emotions.* Norton. [aFG, AR, CT]

Frey, B. & Meier, S. (2004) Social comparison and pro-social behavior: Testing "conditional cooperation" in a field experiment. *American Economic Review* 94:1717–22. Available at: http://www.jstor.org/stable/3592843. [aFG]

Fruteau, C., Voelkl, B., Damme, E. & Noe, R. (2009) Supply and demand determine the market value of food providers in wild vervet monkeys. *Proceedings of the National Academy of Sciences USA* 106:2007–12. [CC]

Fudenberg, D., Levine, D. K. & Maskin, E. (1994) The folk theorem with imperfect public information. *Econometrica* 62:997–1039. Available at: http://www.jstor.org/stable/2951505. [MdS, aFG]

Fudenberg, D. & Maskin, E. (1986) The folk theorem in repeated games with discounting or with incomplete information. *Econometrica* 54:533–54. Available at: http://www.jstor.org/stable/1911307. [aFG]

Fudenberg, D., Rand, D. G. & Dreber, A. (in press) Slow to anger and fast to forgive: Cooperation in an uncertain world. *American Economic Review.* [AD]

Gächter, S. & Herrmann, B. (2009) Reciprocity, culture and human cooperation: Previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364(1518):791–806. Available at: http://rstb.royalsocietypublishing.org/content/364/1518/791.abstract. [AD, SGä, aFG]

Gächter, S. & Herrmann, B. (2011) The limits of self-governance when cooperators get punished: Experimental evidence from urban and rural Russia. *European Economic Review* 55(2):193–210. Available at: http://www.sciencedirect.com/science/article/B6V64-4YYGH46-1/2/773dfeb270ba5da7ea95bb3730ebf3ab. [AD]

Gächter, S., Renner, E. & Sefton, M. (2008) The long-run benefits of punishment. *Science* 322(5907):1510. Available at: http://www.sciencemag.org/content/322/5907/1510 [SGä, arFG]

Gächter, S. & Thöni, C. (2005) Social learning and voluntary cooperation among like-minded people. *Journal of the European Economic Association* 3(2–3):303–14. [SGä]

Gambetta, D. (1993) *The Sicilian Mafia*. Harvard University Press. [rFG]

Gambetta, D. (2009) *Codes of the underworld. How criminals communicate.* Princeton University Press. [WGR]

Gehrig, T., Güth, W., Levati, V., Levinsky, R., Ockenfels, A. & Uske, T. (2007) Buying a pig in a poke: An experimental study of unconditional veto power. *Journal of Economic Psychology* 28(6):692–703. [YB-M]

Gerber, A. S., Green, D. P. & Larimer, C. W. (2008) Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review* 102(1):33–48. Available at: http://dx.doi.org/10.1017/S000305540808009X. [rFG, TJ]

Gigerenzer, G., Todd, P. M. & the ABC Research Group (1999) *Simple heuristics that make us smart*. Oxford University Press. [CC]

Gilby, I. C. (2006) Meat sharing among the Gombe chimpanzees: Harassment and reciprocal exchange. *Animal Behaviour* 71:953–63. [KJ]

Gintis, H. (2000) Strong reciprocity and human sociality. *Journal of Theoretical Biology* 206(2):169–79. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0022519300921118. [AD, HG, aFG, ŞGü]

Gintis, H. (2004) The genetic side of gene-culture coevolution: Internalization of norms and prosocial emotions. *Journal of Economic Behavior and Organization* 53(1):57–67. [HG]

Gintis, H. (2005) Behavioral game theory and contemporary economic theory. *Analyze Kritik* 27(1):48–72. [HG]

Gintis, H. (2006) Behavioral ethics meets natural justice. *Politics, Philosophy and Economics* 5:5–32. Available at: http://ppe.sagepub.com/content/5/1/5.full.pdf + html [aFG]

Gintis, H. (2009) *The bounds of reason: Game theory and the unification of the behavioral sciences.* Princeton University Press. [HG, aFG]

Gintis, H., Bowles, S., Boyd, R. & Fehr, E. (2003) Explaining altruistic behavior in humans. *Evolution and Human Behavior* 24:153–72. [MdS, ŞGü]

Gintis, H., Bowles, S., Boyd, R. & Fehr, E., eds. (2005) *Moral sentiments and material interests.* MIT Press. [RS]

Gintis, H., Boyd, R., Bowles, S. & Fehr, E. (2003) Explaining altruistic behavior in humans. *Evolution and Human Behavior* 24:153–72. Available at: http://dx.doi.org/10.1016/S1090-5138(02)00157-5. [aFG]

Gintis, H., Boyd, R., Bowles, S. & Fehr, E., eds. (2005) *Moral sentiments and material interests: The foundations of cooperation in economic life.* MIT Press. [aFG]

Gintis, H., Smith, E. A. & Bowles, S. (2001) Costly signaling and cooperation. *Journal of Theoretical Biology* 213(1):103–19. [JH]

Glaeser, E. L. & Sacerdote, B. (2000) *The determinants of punishment: Deterrence, incapacitation and vengeance.* Harvard Institute of Economic Research Paper, No. 1894. (SSRN). Available at: http://ssrn.com/paper=236443. [NB]

Gneezy, U. & Rustichini, A. (2000) Pay enough or don't pay at all. *Quarterly Journal of Economics* 115:791–810. [CC]

Goodale, J. (1971) *Tiwi wives: A study of the women of Melville Island, North Australia.* University of Washington Press. [DRe]

Gouldner, A. W. (1960) The norm of reciprocity: A preliminary statement. *American Sociological Review* 25:161–78. Available at: http://www.jstor.org/stable/2092623. [aFG]

Green, E. J. & Porter, R. H. (1984) Noncooperative collusion under imperfect price information. *Econometrica: Journal of the Econometric Society* 52(1) 87–100. [YB-M]

Gros-Louis, J. (2004) The function of food-associated calls in white-faced capuchin monkeys, *Cebus capucinus*, from the perspective of the signaller. *Animal Behaviour* 67:431–40. [KJ]

Guala, F. (2005) *The methodology of experimental economics.* Cambridge University Press. [arFG, ACP]

Guala, F. (2008) Paradigmatic experiments: The ultimatum game from testing to measurement device. *Philosophy of Science* 75:658–69. Available at: http://www.journals.uchicago.edu/doi/abs/10.1086/594512. [arFG, ACP]

Gürerk, O., Irlenbusch, B. & Rockenbach, B. (2006) The competitive advantage of sanctioning institutions. *Science* 312:108–11. Available at: http://www.sciencemag.org/cgi/content/abstract/sci;312/5770/108. [aFG]

Güroğlu, B., van den Bos, W., Rombouts, S. A. R. B. & Crone, E. A. (2010) Unfair? It depends: Neural correlates of fairness in social context. *Social Cognitive and Affective Neuroscience* 4:414–23. [CC]

Gurven, M. (2004) To give and to give not: The behavioral ecology of human food transfers. *Behavioral and Brain Sciences* 27:543–59. Available at: http://journals.cambridge.org/abstract_S0140525X04000123. [NB, aFG]

Gurven, M., Allen-Arave, W., Hill, K. & Hurtado, M. (2000) "It's a Wonderful Life": Signaling generosity among the Ache of Paraguay. *Evolution and Human Behavior* 21:263–82. [CRvR]

Gurven, M. & Winking, J. (2008) Collective action in action: Pro-social behavior in and out of the laboratory. *American Anthropologist* 110(2):179–90. [ACP, CRvR]

Güth, W., Schmittberger, R. & Schwarze, B. (1982) An experimental analysis of ultimatum Bargaining. *Journal of Economic Behavior and Organization* 3(4):367–88. Available at: http://linkinghub.elsevier.com/retrieve/pii/0167268182900117. [YB-M, aFG]

Guzman, R. A., Rodriguez-Sickert, C. & Rowthorn, R. (2007) When in Rome, do as the Romans do: The coevolution of altruistic punishment, conformist learning, and cooperation. *Evolution and Human Behavior* 28(2):112–17. [JH]

Hacking, I. (1988) The participant irrealist at large in the laboratory. *British Journal for the Philosophy of Science* 39:277–94. Available at: http://www.jstor.org/stable/687207. [aFG]

Hagen, E. H. & Hammerstein, P. (2006) Game theory and human evolution: A critique of some recent interpretations of experimental games. *Theoretical Population Biology* 69:339–48. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0040580905001668. [MdS, aFG, AS]

Hamilton, W. D. (1964) The genetical evolution of social behavior. *Journal of Theoretical Biology* 7:1–52. [CRvR]

Harrison, G. W. & List, J. A. (2004) Field experiments. *Journal of Economic Literature* 42:1009–1055. Available at: http://www.ingentaconnect.com/content/aea/jel/2004/00000042/00000004/art00001. [aFG]

Hauert, C., De Monte, S., Hofbauer, J. & Sigmund, K. (2002) Volunteering as a mechanism for cooperation in public goods games. *Science* 296:1129–32. [AD]

Hauert, C., Traulsen, A., Brandt, H., Nowak, M. A. & Sigmund, K. (2007) Via freedom to coercion: The emergence of costly punishment. *Science* 316:1905–907. [AD, CRvR]

Hauser, M. D. (1992) Costs of deception: Cheaters are punished in rhesus monkeys (*Macaca mulatta*). *Proceedings of the National Academy of Sciences USA* 89:12137–39. [KJ]

Haviland, J. B. (1977) *Gossip, reputation, and knowledge in Zinacantan.* University of Chicago Press. [MF]

Hawkes, K. (1993) Why hunter-gatherers work: An ancient version of the problem of public goods. *Current Anthropology* 34:341–62. Available at: http://www.jstor.org/stable/2743748. [aFG]

Henrich, J. (2000) Does culture matter in economic behavior: Ultimatum game bargaining among the Machiguenga. *American Economic Review* 90(4):973–80. [JH]

Henrich, J. (2004) Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior and Organization* 53(1):3–35. [SGä, JH]

Henrich, J. & Boyd, R. (2001) Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology* 208(1):79–89. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0022519300922021. [NB, SB, aFG, JH]

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E. & Gintis, H., eds. (2004) *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies.* Oxford University Press. [aFG, JH]

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., Henrich, N. S., Hill, K., Gil-White, F., Gurven, M., Marlowe, F. W., Patton, J. Q. & Tracer, D. (2005) "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences* 28(6):795–855. [JH, ACP]

Henrich, J., Boyd, R., Bowles, S., Gintis, H., Camerer, C., Fehr, E. & McElreath, R. (2001) In search of Homo economicus: Experiments in 15 small-scale societies. *American Economic Review* 91:73–78. [JH]

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D. P. & Ziker, J. (2010a) Market, religion, community size and the evolution of fairness and punishment. *Science* 327:1480–84. Available at: http://dx.doi.org/10.1126/science.1182238. [HG, rFG, JH]

Henrich, J. & Fehr, E. (2003) Is strong reciprocity a maladaptation? On the evolutionary foundations of human altruism. In: *Genetic and cultural evolution of cooperation*, ed. P. Hammerstein, pp. 55–82. MIT Press. [AS]

Henrich, J., Heine, S. J. & Norenzayan, A. (2010b) Beyond WEIRD: Towards a broad-based behavioral science. *Behavioral and Brain Sciences* 33(2/3):51–75. [JH]

Henrich, J., Heine, S. J. & Norenzayan, A. (2010c) The weirdest people in the world? *Behavioral and Brain Sciences* 33:61–83. Available at: http://dx.doi.org/10.1017/S0140525X0999152X. [rFG]

Henrich, J. & Henrich, N. (2007) *Why humans cooperate: A cultural and evolutionary explanation.* Oxford University Press. [aFG]

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D. & Ziker, J. (2006) Costly punishment across human societies. *Science* 312:1767–70. [KJ]

Herrmann, B., Thöni, C. & Gächter, S. (2008) Antisocial punishment across societies. *Science* 319(5868):1362–67. Available at: http://www.sciencemag.org/cgi/content/abstract/sci;319/5868/1362 [AD, SGä, aFG, JH]

Hertwig, R. & Herzog, S. (2009) Fast and frugal heuristics: Tools of social rationality. *Social Cognition* 27:661–98. [CC]

Hilbe, C. & Sigmund, K. (2010) Incentives and opportunism: From the carrot to the stick. *Proceedings of the Royal Society B: Biological Sciences* 277:2427–33. [MdS, CT]

Hill, K., Barton, M. & Hurtado, A. M. (2009) The emergence of human uniqueness: Characters underlying behavioral modernity. *Evolutionary Anthropology: Issues, News, and Reviews* 18:187–200. [CT]

Hill, K. & Kaplan, H. (1999) Life history traits in humans: Theory and empirical studies. *Annual Review of Anthropology* 28:397–430. [NB]

Hill, K. R., Walker, R., Božičević, M., Eder, J., Headland, T., Hewlett, B., Hurtado, A. M., Marlowe, F., Wiessner, P. & Wood, B. (2011) Coresidence patterns in hunter-gatherer societies show unique human social structure. *Science* 331:1286–89. [CB]

Hirshleifer, J. (1987) On the emotions as guarantors of threats and promises. In: *The latest on the best*, ed. J. Dupré, pp. 307–26. Harvard University Press. [aFG]

Hoebel, E. A. (1954) *The law of primitive man: A study in comparative legal dynamics.* Harvard University Press. [NB]

Hoffman, E., McCabe, K., Shachat, K. & Smith, V. L. (1994) Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior* 7:346–80. [ŞGü]

Hoffman, E., McCabe, K. & Smith, V.L. (1996) Social distance and other-regarding behavior in dictator games.*American Economic Review.* 86, 653–60. [CC]

Hooper, P. L., Kaplan, H. K. & Boone, J. L. (2010) A theory of leadership in human cooperative groups. *Journal of Theoretical Biology* 265(4):633–46. [CRvR]

Howell, P. (1954) *A Manual of Nuer Law: Being an account of customary law, its evolution and development in the courts established by the Sudan government.* International African Institute Publication. Oxford University Press. [NB]

IJzerman, H. & Koole, S. L. (2011) From perceptual rags to metaphoric riches: Bodily, social, and cultural constraints on sociocognitive metaphors. *Psychological Bulletin* 137:355–61. [PAMVL]

Jack, K. B. (2009) Upstream-downstream transactions and watershed externalities: Experimental evidence from Kenya. *Ecological Economics* 68:1813–24. [ACP]

Jacquet, J., Hauert, C., Traulsen, A. & Milinski, M. (2011) Shame and honour drive cooperation. *Biology Letters* 1–3. (Published online before print, June 1, 2011). [ACP]

Janssen, M. A. & Bushman, C. (2008) Evolution of cooperation and altruistic punishment when retaliation is possible. *Journal of Theoretical Biology* 254(3):541–45. Available at: http://www.sciencedirect.com/science/article/B6WMD-4SVC5VF-1/2/9b458b4fcc9244d8f500d6fa35ef6300. [AD]

Janssen, M. A., Holahan, R., Lee, A. & Ostrom, E. (2010) Lab experiments for the study of social–ecological systems. *Science* 328(5978):613–17. Available at: http://www.sciencemag.org/cgi/content/abstract/328/5978/613. [aFG, EO]

Jensen, K. (2010) Punishment and spite, the dark side of cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365:2635–50. [KJ]

Jensen, K. (in press) Social regard: Evolving a psychology of cooperation. In: *The evolution of primate societies*, ed. J. Mitani, J. Call, P. Kappeler, R. Palombit & J. Silk. Chicago University Press. [KJ]

Jensen, K., Call, J. & Tomasello, M. (2007a) Chimpanzees are vengeful but not spiteful. *Proceedings of the National Academy of Sciences USA* 104:13046–50. [KJ]

Jensen, K., Call, J. & Tomasello, M. (2007b) Chimpanzees are rational maximizers in an ultimatum game. *Science* 318:107–109. [KJ]

Jensen, K., Hare, B., Call, J. & Tomasello, M. (2006) What's in it for me? Self-regard precludes altruism and spite in chimpanzees. *Proceedings of the Royal Society of London B: Biological Sciences* 273:1013–21. [KJ]

Jensen, K. & Tomasello, M. (2010) Punishment. In: *Encyclopedia of animal behavior,* ed. M. D. Breed & J. Moore, pp. 800–805. Academic Press. [KJ]

Kahneman, D., Knetsch, J. L. & Thaler, R. H. (1991) Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives* 5:193–206. Available at: http://www.jstor.org/stable/1942711. [aFG]

Kandel, E. & Lazear, E. P. (1992) Peer pressure and partnerships. *Journal of Political Economy*, 100(4):801–17. [MC]

Kelley, H. H., Holmes, J. G., Kerr, N. L., Reis, H. T., Rusbult, C. E. & Van Lange, P. A. M. (2003) *An atlas of interpersonal situations.* Cambridge University Press. [PAMVL]

Kiyonari, T. & Barclay, P. (2008) Free-riding may be thwarted by second-order rewards rather than punishment. *Journal of Personality and Social Psychology* 95(4):826–42. [PB]

Knauft, B. M. (1991) Violence and sociality in human evolution. *Current Anthropology* 32:391–428. Available at: http://www.jstor.org/stable/2743815. [aFG]

Kollock, P. (1993) An eye for an eye leaves everyone blind: Cooperation and accounting systems. *American Sociological Review* 58(6):768–86. [AR]

Konigstein, M. (2000) *Equity, efficiency, and evolutionary stability in bargaining games with joint production.* Springer-Verlag. [CRvR]

Kramer, R. & Cook, K. (2004) *Trust and distrust in organizations: Dilemmas and approaches.* Russell Sage. [GSA]

Krieger, M. J. B. & Ross, K. G. (2002) Identification of a major gene regulating complex social behavior. *Science* 295:328–32. [CC]

Kunreuther, H., Silvasi, G., Bradlow, E. T. & Small, D. (2009) Deterministic and stochastic prisoner's dilemma games: Experiments in interdependent security. *Judgment and Decision Making* 4(5):363–84. [YB-M]

Kurzban, R., DeScioli, P. & O'Brien, E. (2007) Audience effects on moralistic punishment. *Evolution and Human Behavior* 28:75–84. [AS, CT]

Lamba, S. & Mace, R. (2010) People recognize when they are really anonymous in an economic game. *Evolution and Human Behavior* 31:271–78. [ACP]

Langus, A., Petri, J., Nespor, M. & Scharff, C. (in press) *The evolutionary emergence of human language.* Oxford University Press. [CC]

Leaf, M. & Read, D. (in press) *Human thought and social organization: Anthropology on a new plane.* Lexington Books. [DRe]

Ledyard, J. O. (1995) Public Goods: A survey of experimental research, In: *Handbook of experimental economics*, ed. John H. Kagel & Alvin E. Roth, pp. 111–94. Princeton University Press. [AR]

Lee, R. B. (1979) *The !Kung San: Men, women, and work in a foraging society.* Cambridge University Press. [CB, aFG]

Lemonnier, P. (1990) *Guerres et festins. Paix, échanges et compétition dans les Highlands de Nouvelle-Guinée*, Maison de Sciences de L'Homme. [PW]

Lewin, K. (1952) *Field theory in social sciences: Selected theoretical papers*. Harper. [PAMVL]

Lewis, M. (1995) *Shame: The exposed self.* Free Press. [DRo]

List, J. (2006) The behavioralist meets the market: Measuring social preferences and reputation effects in actual transactions. *Journal of Political Economy* 114(1):1–37. [CRvR]

List, J. A. (2009) Social preferences: Some thoughts from the field. *Annual Review of Economics* 1:563–79. [EF]

Lotem, A., Fishman, M. A. & Stone, L. (1999) Evolution of cooperation between individuals. *Nature* 400:226–27. [EF]

MacLean, R. C., Fuentes-Hernandez, A., Greig, D., Hurst, L. D. & Gudelj, I. (2010) A mixture of "cheats" and "co-operators" can enable maximal group benefit. *PLoS Biology* 8(9):e1000486. [EF]

Mahdi, N. Q. (1986) Pukhtunwali: Ostracism and honor among the Pathan Hill tribes. *Ethology and Sociobiology* 7:295–304. Available at: http://linkinghub.elsevier.com/retrieve/pii/0162309586900555. [aFG]

Maier-Rigaud, F. P., Martinsson, P. & Staffiero, G. (2009) Ostracism and the provision of a public good: Experimental evidence. *Journal of Economic Behavior and Organization* 73:387–95. [PvdB]

Malinowski, B. (1926) *Crime and custom in savage society.* Rowman & Littlefield. [NB]

Marlowe, F. W. (2005) Hunter-gatherers and human evolution. *Evolutionary Anthropology* 14:54–67. [ACP]

Marlowe, F. W. (2010) *The Hadza: Hunter-gatherers of Tanzania.* University of California Press. [aFG]

Marlowe, F. W. & Berbesque, J. C. (2008) More "altruistic" punishment in larger societies. *Proceedings of the Royal Society B: Biological Sciences* 275:587–90. [KJ]

Marlowe, F. W., Berbesque, J. C., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Ensminger, J., Gurven, M., Gwako, E., Henrich, J., Henrich, N., Lesorogol, C., McElreath, L. & Tracer, D. (2008) More "altruistic" punishment in larger societies. *Proceedings of the Royal Society B* 275(1634):587–92. Available at: http://rspb.royalsocietypublishing.org/cgi/content/abstract/275/1634/587. [aFG, JH, CRvR]

Marshall, L. (1961) Sharing, talking, giving: Relief of social tensions among the !Kung Bushmen. *Africa* 31:231–49. Available at: http://www.jstor.org/stable/1157263. [aFG]

Marshall, L. (1976) *The !Kung of Nyae Nyae.* Harvard University Press. [DRe]

Masclet, D. (2003) Ostracism in work teams: A public good experiments. *International Journal of Manpower* 24(7):867–87. [PvdB]

Masclet, D., Noussair, C., Tucker, S. & Villeval, M.-C. (2003) Monetary and nonmonetary punishment in the voluntary contributions mechanism. *American Economic Review* 93(1):366–80. Available at: http://www.jstor.org/stable/3132181. [SCä, HG, aFG]

Mathew, S. & Boyd, R. (2011) Punishment sustains large-scale cooperation in prestate warfare. *Proceedings of the National Academy of Sciences USA* 108(28):11375–80. Available at: http://dx.doi.org/10.1073/pnas.1105604108. [SB, rFG]

Mauss, M. (1924/1990) *The gift: The form and reason for exchange in archaic societies*, trans. W. D. Halls. Norton Press. [DRe]

Mauss, M. (1954) *The gift: Forms and functions of exchange in archaic societies.* Cohen & West. [aFG]

McCloskey, D. N. (1972) The enclosure of open fields: Preface to a study of its impact on the efficiency of English agriculture in the eighteenth century. *Journal of Economic History* 32:15–35. Available at: http://www.jstor.org/stable/2117175 [aFG]

Meggitt, M. (1962) *Desert people: A study of the Walbiri Aborigines of Central Australia*. University of Chicago Press. [SB, rFG]

Messick, D. M., Allison, S. T. & Samuelson, C. D. (1988) Framing and communication effects on group members' responses to environmental and social uncertainty. In: *Applied behavioral economics*, ed. S. Maital, pp. 677–700. New York University Press. [YB-M]

Miceli, M. P., Near, J. P. & Dworkin, T. M. (2008) *Whistle-blowing in organizations.* Routledge. [GSA]

Milinski, M., Semmann, D., Bakker, T. C. M. & Krambeck, H.-J. (2001) Cooperation through indirect reciprocity: Image scoring or standing strategy? *Proceedings of the Royal Society of London, Series B: Biological Sciences* 268(1484):2495–501. Available at: http://rspb.royalsocietypublishing.org/content/268/1484/2495.abstract. [AD]

Milinski, M., Semmann, D. & Krambeck, H. J. (2002) Reputation helps solve the "tragedy of the commons." *Nature* 415(6870):424–26. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd = Retrieve&db = PubMed&dopt = Citation&list_uids = 11807552. [AD, AR, CT]

Miller, W. (1990) *Bloodtaking and peacemaking: Feud, law, and society in Saga Iceland.* University of Chicago Press. [NB]

Mitzkewitz, M. & Nagel, R. (1993) Experimental results on ultimatum games with incomplete information. *International Journal of Game Theory* 22(2):171–98. [YB-M]

Mook, D. (1983) In defense of external invalidity. *American Psychologist* 38:379–87. [PB]

Nakamaru, M. & Iwasa, Y. (2005) The evolution of altruism by costly punishment in lattice-structured populations: Score-dependent viability versus score-dependent fertility. *Evolution and Ecological Research* 7(6):853–70. [AD]

Nakamaru, M. & Iwasa, Y. (2006) The coevolution of altruism and punishment: Role of the selfish punisher. *Journal of Theoretical Biology* 240(3):475–88. Available at: http://www.sciencedirect.com/science/article/B6WMD-4HPD5F5-5/2/0002070a3941b038e2b266bf7a91a7ac. [AD]

Nardon, V. (2011) Roncola selvaggia colpisce a Nave S. Felice, *Il Trentino*, April 19, p. 28. [MC]

Near, J. P. & Miceli, M. P. (1995) Effective whistle-blowing. *Academy of Management Review* 20:679–708. [GSA]

Nelissen, R. M. A. (2008) The price you pay: Cost-dependent reputation effects of altruistic punishment. *Evolution and Human Behavior* 29:242–48. [AS, CT]

Nikiforakis, N. (2008) Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics* 92:91–112. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0047272707000643. [AD, aFG]

Nikiforakis, N. & Engelmann, D. (2010) *Altruistic punishment and the threat of feuds.* Department of Economics, University of Melbourne. [aFG]

Nikiforakis, N. & Engelmann, D. (2011) Altruistic punishment and the threat of feuds. *Journal of Economic Behavior and Organization* 78:319–32. [NN]

Nikiforakis, N. & Normann, H. T. (2008) A comparative statics analysis of punishment in public-good experiments. *Experimental Economics* 11:358–69. Available at: http://www.springerlink.com/content/971h6r48182358kw/. [aFG, NN]

Nikiforakis, N., Normann, H. T. & Wallace, B. (2011) Asymmetric enforcement of cooperation in a social dilemma. *Southern Economic Journal* 76:638–59. [CRvR]

Noe, R. & Hammerstein, P. (1994) Biological markets: Supply and demand determine the effect of partner choice in cooperation, mutualism and mating. *Behavioral Ecology and Sociobiology* 35:1–11. [CRvR]

Noussair, C. & Tucker, S. (2005) Combining monetary and social sanctions to promote cooperation. *Economic Inquiry* 43:649–60. Available at: http://dx.doi.org/10.1093/ei/cbi045. [aFG]

Nowak, M. A. & Sigmund, K. (2005) Evolution of indirect reciprocity. *Nature* 437:1291–98. [CT]

O'Carroll, R. E., Foster, C., McGeechan, G., Sandford, K. & Ferguson, E. (2011) The "ick factor," anticipated regret and willingness to become an organ donor. *Health Psychology* 30:236–45. [EF]

Ohtsuki, H., Iwasa, Y., Nowak, M. A. (2009) Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. *Nature* 457:79–82. Available at: http://www.nature.com/nature/journal/vaop/ncurrent/full/nature07601.html. [aFG]

Olson, M. (1965) *The logic of collective action: Public goods and the theory of groups*. Harvard University Press. [RS, CRvR]

Ortony, A., Clore, G. L. & Collins, A. (1988) *The cognitive structure of emotions.* Cambridge University Press. [KJ]

Ostrom, E. (1990) *Governing the commons: The evolution of institutions for collective action.* Cambridge University Press. [MC, aFG, EO, WGR, PvdB]

Ostrom, E. (2000) Collective action and the evolution of social norms. *Journal of Economic Perspectives* 14:137–58. Available at: http://www.jstor.org/stable/2646923. [aFG]

Ostrom, E., Walker, J. & Gardner, R. (1992) Covenants with and without a sword: Self-governance is possible. *American Political Science Review* 86(2):404–17. Available at: http://www.jstor.org/stable/1964229. [arFG, EO, WGR]

Otterbein, K. F. (1986) *The ultimate coercive sanction: A cross-cultural study of capital punishment*. HRAF Press. [CB]

Oxoby, R. J. & Spraggon, J. (2008) Mine and yours: Property rights in dictator games. *Journal of Economic Behavior and Organization* 65:703–13. [ŞGü]

Page, T., Putterman, L. & Unel, B. (2005) Voluntary association in public goods experiments: Reciprocity, mimicry, and efficiency. *Economic Journal* 115:1032–53. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1468–0297.2005.01031.x/full. [aFG]

Paglieri, F. (2007) No more charity, please! Enthymematic parsimony and the pitfall of benevolence. In: *Dissensus and the search for common ground: Proceedings of OSSA 2007,* ed. H. V. Hansen, C. W. Tindale, R. H. Johnson & J. A. Blair, pp. 1–26. OSSA (Ontario Society for the Study of Argumentation), University of Windsor. [CT]

Panchanathan, K. & Boyd, R. (2004) Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature* 432:499–502. [JH, CT]

Panksepp, J. (1998) *Affective neuroscience: The foundations of human and animal emotions.* Oxford University Press. [DRo]

Piazza, J. & Bering, J. M. (2008a) Concerns about reputation via gossip promote generous allocations in an economic game. *Evolution and Human Behavior* 29:172–78. [MF]

Piazza, J. & Bering, J. M. (2008b) The effects of perceived anonymity on altruistic punishment. *Evolutionary Psychology* 6:487–501. [AS]

Piliavin, J. A. & Callero, P. L. (1991) *Giving blood: The development of an altruistic identity.* Johns Hopkins University Press. [EF]

Pillutla, M. M. & Murnighan, J. K. (1996) Unfairness, anger, and spite: Emotional rejections of Ultimatum offers. *Organizational Behavior and Human Decision Processes* 68:208–24. [CC]

Polinsky, A. M. & Shavell, S. (2000) The economic theory of public enforcement of law. *Journal of Economic Literature* 38(1):45–76. [NB]

Posner, R. (1983) *The economics of justice.* Harvard University Press. [NB]

Price, M. E. (2008) The resurrection of group selection as a theory of human cooperation. *Social Justice Research* 21:228–40. [AS]

Putnam, R. (2001) *Bowling alone: The collapse and revival of American community.* Simon & Schuster. [rFG]

Rabin, M. (1993) Incorporating fairness into game theory and economics. *American Economic Review* 83(5):1281–302. [YB-M]

Rand, D. G., Armao, J. J., IV, Nakamaru, M. & Ohtsuki, H. (2010) Anti-social punishment can prevent the co-evolution of punishment and cooperation. *Journal of Theoretical Biology* 265(4):624–32. Available at: http://www.sciencedirect.com/science/article/B6WMD-508K880-2/2/5977afc5e646e284c72d99781f9d19e3. [AD]

Rand, D. G., Dreber, A., Ellingsen, T., Fudenberg, D. & Nowak, M. A. (2009a) Positive interactions promote public cooperation. *Science* 325(5945):1272–75. Available at: http://www.sciencemag.org/cgi/content/abstract/325/5945/1272. [MdS, AD]

Rand, D. G. & Nowak, M. A. (2011) The evolution of anti-social punishment in optional public goods games. *Nature Communications* 2:434. [AD]

Rand, D. G., Ohtsuki, H. & Nowak, M. A. (2009b) Direct reciprocity with costly punishment: Generous tit-for-tat prevails *Journal of Theoretical Biology* 256(1):45–57. [AD]

Rankin, D. J., dos Santos, M. & Wedekind, C. (2009) The evolutionary significance of costly punishment is still to be demonstrated. *Proceedings of the National Academy of Sciences USA* 106:E135–E135. [MdS]

Rapoport, A., Sundali, J. A. & Seale, D. A. (1996) Ultimatums in two-person bargaining with one-sided uncertainty: Demand games. *Journal of Economic Behavior and Organization* 30(2):173–96. [YB-M]

Read, D. (2001) What is kinship? In: *The cultural analysis of kinship: The legacy of David Schneider and its implications for anthropological relativism*, ed. R. Feinberg & M. Ottenheimer, pp. 78–117. University of Illinois Press. [DRe]

Read, D. (2007) Kinship theory: A paradigm shift. *Ethnology* 46:329–64. [DRe]

Read, D. (2008) An interaction model for resource implement complexity based on risk and number of annual moves. *American Antiquity* 73:599–625. [DRe]

Read, D. (2010a) Agent-based and multi-agent simulations: Coming of age or in search of an identity? *Computational and Mathematical Organization Theory* 16:329–47. [DRe]

Read, D. (2010b) From experiential-based to relational-based forms of social organization: A major transition in the evolution of *Homo sapiens.* In: *Social brain, distributed mind*, ed. R. Dunbar, C. Gamble & J. Gowlett, pp. 199–230. Oxford University Press. [DRe]

Read, D. (2012) *How culture makes us human: Primate kinship evolution and the formation of human societies.* Left Coast Press. [DRe]

Reuben, E. & Suetens, S. (2011) Revisiting strategic versus non-strategic cooperation. *CentER Discussion Paper 2009–2022.* Tilburg University. [NN]

Richerson, P. & Boyd, R. (1998) The evolution of ultrasociality. In: *Indoctrinability, ideology and warfare*, ed. I. Eibl-Eibesfeldt & F. K. Salter, pp. 71–95. Berghahn Books. [JH]

Richerson, P. J. & Boyd, R. (2005) *Not by genes alone: How culture transformed human evolution.* University of Chicago Press. [aFG, CT]

Robinson, G. E., Fernald, R. D. & Clayton, D. F. (2008) Genes and social behavior. *Science* 322:896–900. [CC]

Rockenbach, B. & Milinski, M. (2006) The efficient interaction of indirect reciprocity and costly punishment. *Nature* 444(7120):718–23. Available at: http://www.nature.com/nature/journal/vaop/ncurrent/full/nature05229. [MdS, AD, MF, SGä, aFG, ACP]

Rosas, A. (2008) The return of reciprocity: A psychological approach to the evolution of cooperation. *Biology & Philosophy* 24:555–66. Available at: http://www.springerlink.com/index/p93381m4744v4g20.pdf. [aFG]

Ross, D. (2006) Evolutionary game theory and the normative theory of institutional design: Binmore and behavioral economics. *Politics, Philosophy, and Economics* 5:51–79. Available at: http://ppe.sagepub.com/content/5/1/51. [aFG]

Ross, D. (2007) *H sapiens* as ecologically special: What does language contribute? *Language Sciences* 29:710–31. [DRo]

Ross, L. (1977) The intuitive psychologist and his shortcomings: Distortions in the attribution process. In: *Advances in experimental social psychology, vol. 10*, ed. L. Berkowitz, pp. 173–220. Academic Press. [GSA]

Roth, A., Prasnikar, V., Okuno-Fujiwara, M. & Zamir, S. (1991) Bargaining and market behavior in Jerusalem, Lubljana, Pittsburgh and Tokyo: An experimental study. *American Economic Review* 81:1068–95. Available at; http://www.jstor.org/stable/2006907. [aFG]

Rothschild, J. & Miethe, T. D. (1999) Whistle-blower disclosures and management retaliation: The battle to control information about organization corruption. *Work and Occupations* 26:107–28. [GSA]

Russell, Y. I., Call, J. & Dunbar, R. I. M. (2008) Image scoring in great apes. *Behavioural Processes* 78:108–11. [CT]

Rustagi, D., Engel, S. & Kosfeld, M. (2010) Conditional cooperation and costly monitoring explain success in forest commons management. *Science* 330(6006):961–65. Available at: http://www.sciencemag.org/content/330/6006/961 [SB, SGä, arFG, JH, ACP]

Ruttan, L. (2008) Economic heterogeneity and the commons: Effects on collective action and collective goods provisioning. *World Development* 36(5):969–85. [CRvR]

Sahlins, M. (1972/1974) *Stone Age economics.* Aldine Transaction/Routledge. (Routledge edition, 1974, cited in Guala T.A.). [aFG, DRe]

Sainty, B. (1999) Achieving greater cooperation in a noisy prisoner's dilemma: An experimental investigation. *Journal of Economic Behavior and Organization* 39(4):421–35. [YB-M]

Sanchez, C. (2005) U.S. Government punishes schools that ban military recruiting. *NPR News* (Online), June 1, 2005. Available at: http://www.npr.org/templates/story/story.php?storyId = 4675926. [GSA]

Sanfey, A. G., Rilling, J. K., Aaronson, J. A., Nystrom, L. E. & Cohen, J. D. (2003) The neural basis of economic decision-making in the Ultimatum Game. *Science* 300:1755–58. Available at: http://www.sciencemag.org/cgi/content/abstract/300/5626/1755. [CC, aFG]

Santos, M. D., Rankin, D. J. & Wedekind, C. (2011) The evolution of punishment through reputation. *Proceedings of the Royal Society B: Biological Sciences* 278:371–77. [AS]

Schapera, I. (1956) *Government and politics in tribal societies.* Watts. [WGR]

Semmann, D., Krambeck, H.-J. & Milinski, M. (2004) Strategic investment in reputation. *Behavioral Ecology and Sociobiology* 56:248–52. [CT]

Shang, J. & Croson, R. (2009) A field experiment in charitable contribution: The impact of social information on the voluntary provision of public goods. *Economic Journal* 119:1422–39. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1468–0297.2009.02267.x/full. [aFG]

Sigmund, K. (2007) Punish or perish? Retaliation and collaboration among humans. *Trends in Ecology and Evolution* 22:593–600. [MdS]

Sigmund, K., De Silva, H., Traulsen, A. & Hauert, C. (2010) Social learning promotes institutions for governing the commons. *Nature* 466(7308):861–63. Available at: (1) http://dx.doi.org/10.1038/nature09203; (2) http://www.nature.com/nature/journal/v466/n7308/abs/nature09203.html#supplementary-information. [AD]

Silk, J. B. & House, B. R. (2011) Evolutionary foundations of human prosocial sentiments. *Proceedings of the National Academy of Sciences USA* 108 (Suppl. 2):10910–17. [KJ]

Smirnov, O., Dawes, C. T., Fowler, J. H., Johnson, T. & McElreath, R. (2010) The behavioral logic of collective action: Partisans cooperate and punish more than non-partisans. *Political Psychology* 31(4):595–616. [TJ]

Smith, E. A., Bleige-Bird, R. & Bird, D. W. (2003) The benefits of costly signaling: Meriam turtle hunters. *Behavioral Ecology* 14(1):116–26. [JH, PW]

Smith, V. L. (1982) Microeconomic systems as an experimental science. *American Economic Review* 72(5):923–55. [SGä]

Sober, E. & Wilson, D. S. (1998) *Unto others: The evolution and psychology of unselfish behavior.* Harvard University Press. [arFG]

Sommerfeld, R. D., Krambeck, H., Semmann, D. & Milinski, M. (2007) Gossip as an alternative for direct observation in games of indirect reciprocity. *Proceedings of the National Academy of Sciences USA* 104:17435–40. [MF]

Spitzer, M., Fischbacher, U., Herrnberger, B., Grön, G. & Fehr, E. (2007) The neural signature of norm compliance. *Neuron* 56:185–96. Available at: http://linkinghub.elsevier.com/retrieve/pii/S089662730700709X. [EF, aFG]

Stanton, M. (2000) *From Selma to sorrow: The life and death of Viola Liuzzo.* University of Georgia Press. [GSA]

Starmer, C. (1999) Experiments in economics ... (should we trust the dismal scientists in white coats?). *Journal of Economic Methodology* 6:1–30. Available at: http://dx.doi.org/10.1080/135017899000000001. [aFG]

Steel, D. (2007) *Across the boundaries: Extrapolation in biology and in the social sciences.* Oxford University Press. [aFG]

Strathern, A. (1971) *The rope of Moka: Big-men and ceremonial exchange in Mt. Hagen, Papua New Guinea.* Cambridge University Press. [PW]

Strauss, S. Y., Rudgers, J. A., Lau, J. A. & Irwin, R. E. (2002) Direct and ecological costs of resistance to herbivory. *Trends in Ecology and Evolution* 17(6):278–85. [PvdB]

Strehlow, T. G. H. (1970) Geography and the totemic landscape in Central Australia: A functional study. In: *Australian aboriginal anthology: Modern studies in the social anthropology of the Australian aborigines*, ed. R. M. Berndt, pp. 92–140. University of Western Australia Press. [SB, rFG]

Sugden, R. (1984) Reciprocity: The supply of public goods through voluntary contributions. *Economic Journal* 94:772–87. [RS]

Sugden, R. (2007) Review of: *Moral sentiments and material interests*, ed. H. Gintis, S. Bowles, R. Boyd & E. Fehr (2005). *Economica* 74:371–78. [RS]

Sunstein, C., Schkade, D. & Kahneman, D. (2000) Do people want optimal deterrence? *Journal of Legal Studies* 29(1):237–53. [NB]

Sylwester, K. & Roberts, G. (2010) Cooperators benefit through reputation-based partner choice in economic games. *Biology Letters* 6:659–62. [CT]

Tennie, C., Frith, U. & Frith, C. D. (2010) Reputation management in the age of the world-wide web. *Trends in Cognitive Sciences* 14:482–88, Available at; http://www.sciencedirect.com/science/article/pii/S136466131000149X. [CT]

Thorndike, E. L. (1927) The law of effect. *American Journal of Psychology* 39:212–22. [AS]

Tinbergen, N. (1963) On aims and methods of ethology. *Zeitschrift für Tierpsychologie* 20:410–33. [MdS]

Tinbergen, N. (1968) On war and peace in animals and man. *Science* 160:1411–18. [PB]

Tom, S. M., Fox, C. R., Trepel, C. & Poldrack, R. A. (2007) The neural basis of loss aversion in decision-making under risk. *Science* 315:515–18. Available at: http://www.sciencemag.org/cgi/content/abstract/315/5811/515. [aFG]

Tooby, J., Krasnow, M., Delton, A. & Cosmides, L. (2009) I will only know that our interaction was one-shot if I kill you: A cue theoretic approach to the architecture of cooperation. Talk presented at the Human Behavior and Evolution Society Meeting, California State University, Fullerton, CA, May 2009. [AS]

Torrence, R. (1989) Re-tooling: Towards a behavioral theory of stone tools, In: *Time, energy and stone tools*, ed. R. Torrence, pp. 57–66. Cambridge University Press. [DRe]

Traulsen, A., Hauert, C., De Silva, H., Nowak, M. A. & Sigmund, K. (2009) Exploration dynamics in evolutionary games. *Proceedings of the National Academy of Sciences USA* 106(3):709–12. Available at: http://www.pnas.org/content/106/3/709.abstract [AD]

Trevino, L. K. (1992) The social effects of punishment in organizations: A justice perspective. *Academy of Management Review* 17(4):647–76. [GSA]

Trivers, R. L. (1971) The evolution of reciprocal altruism. *Quarterly Review of Biology* 46:35–57. Available at: http://www.jstor.org/stable/2822435. [NB, MdS, aFG, AR]

Trivers, R. L. (1972) Parental investment and sexual selection. In: *Sexual selection and the descent of man*, ed. B. G. Campbell, pp. 136–207. Aldine. [aFG]

Trivers, R. L. (2004) Mutual benefits at all levels of life. [review of *Genetic and cultural evolution of cooperation*, Peter Hammerstein, Ed.] *Science* 304:964–65. [aFG, CT]

Trompf, G. W. (1994) *Payback: The logic of retribution in Melanesian religions.* Cambridge University Press. [PW]

Turnbull, C. (1961) *The forest people.* Jonathan Cape. [aFG]

Turner, M. M., Mazur, M. A., Wendel, N. & Winslow, R. (2003) Relational ruin or social glue? The joint effect of relationship type and gossip valence on liking, trust, and expertise. *Communication Monographs* 70:129–41. [GSA]

Ule, A., Schram, A., Riedl, A. & Cason, T. N. (2009) Indirect punishment and generosity towards strangers. *Science* 326(5960):1701–704. Available at: http://www.sciencemag.org/cgi/content/abstract/326/5960/1701. [EF, SGä, aFG]

van den Steenhoven, G. (1962) *Leadership and law among the Eskimos of the Keewatin District, Northwest Territories.* Rijswijk Excelsior. [CB]

van der Heijden, E., Potters, J. & Sefton, M. (2009) Hierarchy and opportunism in teams. *Journal of Economic Behavior and Organization* 69(1):39–50. [CRvR]

Van Lange, P. A. M., Klapwijk, A. & Van Munster, L. (2011) How the shadow of the future might promote cooperation. *Group Processes and Intergroup Relations*. 14:857–70. [PAMVL]

Van Lange, P. A. M., Ouwerkerk, J. W. & Tazelaar, M. J. A. (2002) How to overcome the detrimental effects of noise in social interaction: The benefits of generosity. *Journal of Personality and Social Psychology* 82:768–80. [PAMVL]

Van Lange, P. A. M. & Rusbult, C. E. (2012) Interdependence theory. In: *Handbook of theories of social psychology, vol. 2*, ed. P. A. M. Van Lange, A. W. Kruglanski & E. T. Higgins, pp. 251–72. Sage. [PAMVL]

van't Wout, M., Kahn, R. S., Sanfey, A. G. & Aleman, A. (2006) Affective state and decision-making in the Ultimatum Game. *Experimental Brain Research* 169:564–68. [CC]

von Führer-Haimendorf, C. (1967) *Morals and merit: A study of values and social control in South-Asian societies.* Weidenfeld & Nicolson. [NB]

von Rueden, C., Gurven, M. & Kaplan, H. K. (2010) Face-to-face collective action games in a small-scale society: Testing the effects of leadership on group performance and the division of spoils. Paper presented at the Initiative for the Study of Religion, Economics, and Society, Chapman University. Orange, CA, November 12, 2010. [CRvR]

von Rueden, C., Hooper, P. L., Gurven, M. & Kaplan, H. K. (2009) The patterning of male conflict with kinship, cooperative networks, and social status in a small-scale Amazonian society. Paper presented at the Annual Meeting of the Human Behavior and Evolution Society, Fullerton, CA, May 30, 2009. [CRvR]

Vyrastekova, J. & van Soest, D. (2008) On the (in)effectiveness of rewards in sustaining cooperation. *Experimental Economics* 11(1):53–65. Available at: http://dx.doi.org/10.1007/s10683-006-9153-x. [AD]

Walker, J. M. & Halloran, M. A. (2004) Rewards and sanctions and the provision of public goods in one-shot settings. *Experimental Economics* 7:235–47. Available at: http://www.springerlink.com/content/u72765n343004k60/. [aFG]

Wang, J., Wu, B., W. C. Ho, D. & Wang, L. (2011) Evolution of cooperation in multilevel public goods games with community structures. *European Physical Letters* 93(5):58001. Available at: http://dx.doi.org/10.1209/0295-5075/93/58001. [AD]

Wedekind, C. & Braithwaite, V. A. (2002) The long-term benefits of human generosity in indirect reciprocity. *Current Biology* 12:1012–15. [CT]

Wedekind, C. & Milinski, M. (2000) Cooperation through image scoring in humans. *Science* 288(5467):850–52. Available at: http://www.sciencemag.org/cgi/content/abstract/288/5467/850. [AD, CT]

Weingast, B. R. & Moran, M. J. (1983) Bureaucratic discretion or congressional control? Regulatory policymaking by the Federal Trade Commission. *Journal of Political Economy* 91(5):765–800. [TJ]

West, S. A., El Mouden, C. & Gardner, A. (2011) Sixteen common misconceptions about the evolution of cooperation in humans. *Evolution and Human Behavior* 32(4):231–62. [PB, MdS, rFG, ACP]

West, S. A., Griffin, A. S. & Gardner, A. (2007a) Evolutionary explanations for cooperation. *Current Biology* 17:R661–72. [MdS]

West, S. A., Griffin, A. S. & Gardner, A. (2007b) Social semantics: Altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology* 20:415–32. [PB]

Wiessner, P. (1977) Hxaro, a regional system of reciprocity for reducing risk among the !Kung San. Unpublished PhD dissertation, University of Michigan. [DRe]

Wiessner, P. (1982) Risk, reciprocity and social influences on !Kung San economies. In: *Politics and history in band societies*, ed. H. R. Leacock & R. B. Lee, pp. 61–84. Cambridge University Press. [DRe]

Wiessner, P. (2002) Hunting, healing, and Hxaro exchange: A long term perspective on !Kung (Ju/'hoansi) large-game hunting. *Evolution and Human Behavior* 23:1–30. [PW]

Wiessner, P. (2005) Norm enforcement among the Ju/'hoansi bushmen: A case for strong reciprocity? *Human Nature* 16(2):115–45. Available at: http://www.springerlink.com/index/dg3m0660x4lgdl9t.pdf. [MC, HG, arFG, CRvR, PW]

Wiessner, P. (2009) Experimental games and games of life among the Ju/'hoan Bushmen. *Current Anthropology* 50(1):133–38. Available at: http://www.jstor.org/stable/20479691. [HG, arFG, DRe, CRvR, PW]

Wiessner, P. & Tumu, A. (1998) *Historical vines: Enga networks of exchange, ritual and warfare in Papua New Guinea.* Smithsonian Institution Press. [PW]

Willer, R. (2009) Groups reward individual sacrifice: The status solution to the collective action problem. *American Sociological Review* 74:23–43. [MF]

Willer, R., Feinberg, M., Irwin, K., Schultz, M. & Simpson, B. (2010) The trouble with invisible men: How reputational concerns motivate generosity. In: *The handbook of sociology of morality*, ed. S. Hitlin & S. Vaisey, pp. 315–30. Springer. [MF]

Wilson, D. S. (1979) Structured demes and trait-group variation. *American Naturalist* 113:606–10. Available at: http://www.jstor.org/stable/2460279. [aFG]

Wilson, D. S. & Sober, E. (1994) Reintroducing group selection to the human behavioral sciences. *Behavioral and Brain Sciences* 17:585–654. Available at: http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=6750548&fulltextType=RA&fileId=S0140525X00036104 [aFG]

Wischniewski, J., Windmann, S., Juckel, G. & Brune, M. (2009) Rules of social exchange: Game theory, individual differences and psychopathology. *Neuroscience and Biobehavioral Reviews* 33:305–13. [EF]

Wolf, M., van Doorn, G. S., Leimar, O. & Weissing, F. J. (2007) Life-history trade-offs favour the evolution of animal personalities. *Nature* 447:581–84. [PvdB]

Wolf, M., van Doorn, G. S. & Weissing, F. J. (2008) Evolutionary emergence of responsive and unresponsive personalities. *Proceedings of the National Academy of Sciences USA* 105(41):15825–30. [PvdB]

Woodburn, J. (1982) Egalitarian societies. *Man* 17:431–51. [CB]

Wu, J.-J., Zhang, B.-Y., Zhou, Z.-X., He, Q.-Q., Zheng, X.-D., Cressman, R. & Tao, Y. (2009) Costly punishment does not always increase cooperation. *Proceedings of the National Academy of Sciences USA* 106(41):17448–451. Available at: http://www.pnas.org/content/106/41/17448.abstract. [MdS, AD]

Xiao, E. & Houser, D. (2005) Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences* 102:7398–401. Available at: http://www.pnas.org/content/102/20/7398.full. [aFG]

Yacubian, J., Gläscher, J., Schroeder, K., Sommer, T., Braus, D,F. & Büchel, C. (2006) Dissociable systems for gain- and loss-related value predictions and errors of prediction in the human brain. *Journal of Neuroscience* 26:9530–37. Available at: http://www.jneurosci.org/cgi/content/abstract/26/37/9530. [aFG]

Yamagishi, T. (1986) The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology* 51:110–16. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0022351407601885. [arFG]

Zizzo, D. J. & Oswald, A. J. (2001) Are people willing to pay to reduce others' incomes? *Annales d'économie et de statistique* 63–64:39–65. [KJ]