RECIPROCITY: WEAK OR STRONG?

WHAT PUNISHMENT EXPERIMENTS DO (AND DO NOT) DEMONSTRATE

FRANCESCO GUALA

Working Paper n. 2010-23

LUGLIO 2010

DIPARTIMENTO DI SCIENZE ECONOMICHE AZIENDALI E STATISTICHE

Via Conservatorio 7
20122 Milano
tel. ++39 02 503 21501 (21522) - fax ++39 02 503 21450 (21505)
http://www.economia.unimi.it
E Mail: dipeco@unimi.it

Reciprocity: Weak or Strong?
What Punishment Experiments Do (and Do Not) Demonstrate*

Francesco Guala†

Abstract

Strong Reciprocity theorists claim that cooperation in social dilemma games can be sustained by costly punishment mechanisms that eliminate incentives to free ride, even in one-shot and finitely repeated games. There is little doubt that costly punishment raises cooperation in laboratory conditions. Its efficacy in the field however is controversial. I distinguish two interpretations of experimental results, and show that the wide interpretation endorsed by Strong Reciprocity theorists is unsupported by ethnographic evidence on decentralised punishment and by historical evidence on common pool institutions. The institutions that spontaneously evolve to solve dilemmas of cooperation typically exploit low-cost mechanisms, turning finite games into indefinitely repeated ones and eliminating the cost of sanctioning.

JEL Classification: D02, D03, C92, H41, Z1

Keywords: Experiments; Cooperation; Punishment; Evolution.

* Previous versions of this paper were presented at Bocconi University, the Max Planck Institute for Economics in Jena, and STOREP 2010. I was helped during revision by Paolo Garella, Herbert Gintis, Alessandro Innocenti, Josh Miller, Ivan Moscati, Elinor Ostrom, Nikos Nikiforakis, Alejandro Rosas, Don Ross, and Jim Woodward's generous comments. All the remaining mistakes are mine.

† Department of Economics, Business and Statistics, University of Milan, via Conservatorio 7, 20122 Milan, Italy. Email: francesco.guala@unimi.it

1. Introduction

Over the last two decades research on human cooperation has made considerable progress both on the theoretical and the empirical front. Economists and biologists have proposed a distinction between two kinds of mechanisms – "Strong" and "Weak" Reciprocity – that may explain the evolution of human sociality. Reciprocity is, broadly speaking, a tendency to respond "nice" to "nice" 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. Or, to put it differently: a weak reciprocator will choose only strategies that are part of a Nash-equilibrium of the game she is playing.

Strong Reciprocity models, in contrast, allow players to choose sub-optimal strategies, and thus diverge substantially from the models of self-interested behaviour that are typically used by evolutionary biologists and rational choice theorists. The behaviour of strong reciprocators can be less than optimal in roughly two ways: on the one hand, a strong reciprocator plays cooperatively with cooperators, even though it would be more advantageous to exploit them (let us call it *positive* Strong Reciprocity). On the other, a strong reciprocator is willing to punish defectors at a cost for herself, 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, even a small fraction of free riders can drive positive reciprocators towards low levels of cooperation in finitely repeated games. Costly punishment in such circumstances may provide just 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. On deeper inspection, however, the message of these experiments is far from clear. To dispel some confusion, it will be helpful to introduce some preliminary distinctions between concepts that are often conflated in the writings of reciprocity theorists. It will turn out that some 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 paper is to clarify the concepts used by economists and biologists and help the resolution of open issues in reciprocity theory.

I will 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. Secondly, studies by economic and social historians show that social dilemmas in the wild are typically solved by institutions that eliminate the costs of decentralized punishment and facilitate the application of non-costly punishment mechanisms. The second goal of this paper then is to survey relevant evidence from history and anthropology that economists interested in reciprocity theory are usually unfamiliar with.

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 what 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 (Figure 1).[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

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.

[1] 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.

"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 (Trivers 1971, Dawkins 1976, Axelrod and Hamilton 1981). 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" or "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. Reciprocity can sustain cooperation in the long run, and pairs of reciprocators are more efficient producers of resources than selfish free riders.[3]

Axelrod's 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, we shall classify

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

[3] While Axelrod's results have not survived rigorous scrutiny, some of his insights have been shown to hold in certain classes of models. Interested readers are referred to Bendor and Swistak (1995, 1997) and Binmore (1998, Ch. 3).

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" (I shall present them in more detail in the next section).

Axelrod's and Trivers' findings are consistent with a general game-theoretic result known as the *folk theorem*.[4] Informally, the folk theorem says that any strategy guaranteeing more than 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. 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". In the matrix of Figure 1, such a profile delivers an average payoff of 2.14 to me, and 1.85 to you. Because it is better than the worst penalty I can inflict (by withdrawing cooperation) if you don't follow it, it is a Nash equilibrium in 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 processes drive 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 (Boyd and Richerson 1990, Wilson and Sober 1994, Bergstrom 2002). If homogenous groups of conditional cooperators are more efficient, in the long run they should be able to outperform homogenous

[4] Fudenberg and Maskin (1986), Fudenberg, Levin and Maskin (1994).

groups of free riders trapped in sub-optimal 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 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' "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.[5] Second, the number of players must be small, so that monitoring cooperation is relatively easy and costless. 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 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

[5] In an objective *and* subjective sense: the players must not discount future payoffs too heavily. If they do, the temptation to defect will be strong regardless of the future stream of gains from cooperation.

best-known advocates, but many other social scientists and biologists have contributed to its success.[6] 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. Being targeted towards single individuals, costly punishment is not indiscriminate like the trigger strategies of standard game theory. 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. This 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 benefits the whole group, but imposes a cost on the punisher only. In principle everybody would like free riders to be punished, but would prefer somebody else to do it. The second-order dilemma can be 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. But could these mechanisms have survived Darwinian selection?

Simulations suggest that cooperative strategies can evolve in favourable conditions (Boyd and Richerson 1992, Gintis 2000, Henrich and Boyd 2001, Boyd et al. 2003, Bowles and Gintis 2003). 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 the more

[6] See e.g. the papers in Henrich et al (eds. 2004), Gintis et al (eds. 2005). Seminal papers are Yamagishi (1986), Boyd and Richerson (1992), and Ostrom et al (1992). Given the plurality of contributions, I can only try to capture an ideal-type of Strong Reciprocity theory, or a hard core of ideas shared by most (but not all) supporters. While some theorists for example consider Strong Reciprocity the key mechanism sustaining human cooperation, others – such as Gintis – view it as just one element in a complex set of mechanisms including character virtues and internalized norms (see e.g. Gintis 2009).

cooperative groups. The problem of multiple equilibria is made more severe by strong reciprocators (Boyd and Richerson 1992) because Strong Reciprocity can support an even wider range of equilibria, including equilibria that are not welfare- or fitness-enhancing for a group. A community for example may be prevented from adopting a set of strategies that would be beneficial for the group, simply because they depart from what is considered "correct" behaviour by an aggressive group 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 experiment, 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 that will be the main focus of this paper.

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 of sharing a sum of money, x. The disagreement point (or status quo) is symmetric – zero for both players – but 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. The unique sub-game perfect equilibrium of the Ultimatum game is for Player 1 to offer $(x - e, \ e)$, where e is the smallest positive unit of surplus division. Player 2 then must accept because e 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 cake, 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 fulfil a useful social

function, for unfair players learn what is expected of them, and conform to the prevailing norm in future encounters.[7]

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.[8] 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.

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

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

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

Tables 2a, b, c

Adding the punishment phase radically changes the game. Take the simplest case of a two-player Prisoner's Dilemma: the matrix in Table 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 Table 2b.[9] 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

[7] There is evidence that the notion of "fair offer" varies across cultures. While 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).

[8] E.g. Fehr and Gachter (2000), Fischbacher et al. (2001), Falk et al. (2003), Fehr and Fischbacher (2004, 2005). Yamagishi (1986) and Ostrom et al (1992) pioneered this approach.

[9] I have used a non-standard matrix for presentational ease. The PP cell in Table 3 is empty to indicate that sanctioned individuals typically do not have the possibility 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 cross-cultural studies of anti-social punishment (e.g. Herrmann et al 2008) but we shall ignore it for the time being.

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 Table 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 and Gachter 2000, 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 effective even when it is administered by "bystanders" or "third parties" – i.e. when the potential punishers are not themselves the victims but have merely witnessed exploitative behaviour (Fehr and Fischbacher 2004).

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 2002) and partly motivated by the sheer pleasure of punishing social deviants (deQuervein et al 2004).[10] 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 questions, various models have been proposed in the game theory literature, and probably even more will appear in the future (see Falk and Fischbacher (2005) for a survey).

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 the Punishment experiments – "narrow" and "wide" – which have rather 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

[10] This evidence seems to confirm the old insight that emotions can help us in situations where rational deliberation delivers sub-optimal outcomes (Hirsheifler 1987, Frank 1988).

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 and Fehr 2004: 55). The punishment design seems to be motivated 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 and Schmidt 2006: 621)

Under this interpretation, punishment mechanisms are useful *methodological devices to observe social preferences*.[11] 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 typically 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 Gachter claim that "in our view punishment of free-riding also plays an important role in real life" (2000: 993). Influential anthropologists Boyd and Richerson add that

[11] 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 also Guala (2008).

> Fehr's experiment suggests that some of the neighbours 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 and Boyd 2005: 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 and Fehr 2004: 56)

> 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 and Fischbacher 2005: 169; see also Camerer and Fehr 2004: 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 and economics. 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).[12] The wide interpretation makes the additional claim that experimental results can be extrapolated to

[12] This is not a special problem of social science experiments, to be sure: the issue of external validity arises in all sciences, including medicine, biology and physics. But the sheer complexity of social mechanisms is likely to make the extrapolation of experimental results a more delicate matter than in other disciplines. See Guala (2005), Steel (2007), and Bardsley et al (2009) for a thorough discussion of these methodological issues.

explain cooperation in some non-laboratory conditions, and so amounts to an *external validity* inference that requires extra evidence to be sustained.

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 camp 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 centralised policing) to prevent exploitation, free riding, and abuse of power (e.g. Binmore 2005: 82). Weak Reciprocity theory, in other words, draws different boundaries for spontaneous cooperation, and cannot be blamed (without begging the question) for its "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 been observed across subject pools in several developed countries, as well as in Ultimatum and Public Goods experiments ran in small-scale societies (Henrich et al 2004, Marlowe et al 2008, Herrmann 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 field 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. Experiments with punishment in the field are artefactual in this sense, for they involve situations that are probably quite unfamiliar to the decision makers, and as we shall see do not reproduce the full menu of strategies that are available in the dilemmas of cooperation that people face in everyday life. It is more appropriate to speak of "experiments in the field" in this case, rather than "field experiments" proper.

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 either that experimental phenomena such as costly punishment are somewhat "unreal". A phenomenon may be real *and* artefactual – a real experimental effect generated in circumstances that do not mirror those naturally found in our societies and markets.[13] As we shall see, there are good reasons to believe that costly punishment is a "real artefact" in this sense of the term: artefactual in so far 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.

[13] The concept of "artefact" is used in different ways in experimental science. The artefacts reported in microscopy textbooks, for example, can be divided in two categories: those that appear to be, but are not real; and those that are real, but not natural (Hacking 1988). Fringes caused by optical aberrations around the edges of a cell belong to the first category of artefacts. It is the microscope that generates the illusion of fringes, which in fact are not really there. Bubbles on a slide, stains, scratches, folds produced during the preparation of the assay belong in contrast to a second category of artefacts. They are real, but they are produced by the experimental procedure. For instance, if the membranal border of an organelle seems to be interrupted, this may be due to the chemicals used to preserve the tissue. The "natural" membrane was continuous, but the chemical substances used by the experimenter caused its deterioration; not being aware of this fact, the experimenter might infer that it was a characteristic of the cell independently any laboratory manipulation.

7. Repetition and evolutionary scale

Understandably, external validity is not a very popular concept among experimental economists. External validity objections can hinder scientific progress when they are meant to raise sceptical doubts about the use of experiments generally (cf. Starmer 1999, Guala 2005, Bardsley et al 2009). But external validity worries are inescapable and indeed useful when addressed to the specific details of an experimental design, for in such cases they help establish the reliability of specific inferences from the laboratory to field settings. 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 heuristic rules 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 worked well (and was selected for) in the larger but more familiar games that we play in real life. If these games were repeated long enough, however, out-of-equilibrium strategies would be eliminated by evolutionary forces and learning, until behaviour approaches the rational equilibria of the games we play (cf. Binmore 1998, 1999, 2006, Trivers 2004, Hagen and Hammerstein 2006).

This argument sounds plausible, but unfortunately is inconclusive. To begin with, it is easy to retort that pro-social behaviour in settings such as the Ultimatum Game is remarkably robust even when the games are repeated for several rounds (e.g. Roth et al. 1991, Cooper and Dutcher 2009). If Strong Reciprocity "misfires" in finitely repeated games, it does so systematically enough to be of theoretical interest for social scientists (Boyd and Richerson 2005: 220), because 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. Second, there is evidence that experimental subjects distinguish between one-shot and finitely repeated games, and modulate their strategies accordingly. To insist that they do not understand the

difference between finitely and indefinitely repeated games, therefore, seems arbitrary and unjustified (Gintis et al. 2003, Fehr and Fischbacher 2002, 2005).

These replies are powerful and the critics of Strong Reciprocity theory are wrong to insist with 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 costly 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 two sections review non-experimental evidence that is seldom discussed by theorists on either side of the controversy. We shall begin with ethnographic data from anthropology (sections 8-11), a source that is often cited by reciprocity theorists but never analysed in much depth. Section 12 deals instead with historical evidence concerning common pool institutions.

8. Costly punishment in small societies

The Leviathan is a relatively recent invention. During most of its 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 1974: 186-7)

The small-scale societies of hunter-gatherers, horticolturalists, 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. In their writings, Strong Reciprocity theorists say that their models explain the emergence and maintenance of cooperation in small egalitarian societies, but provide surprisingly thin evidence in support.

According to Bowles and Gintis (2002: 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 refer to a study (Boehm 1999) that does *not* endorse a costly punishment account of human sociality. Richerson and Boyd (2005: 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 and Boyd 2005: 280, n. 60).

Most of Richerson and Boyd's case, in fact, is based on Fehr and Gachter's (2000, 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: 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 and Gachter 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 (Sober and Wilson 1998: 166-168).

The costly punishment account of cooperation in small societies, then, seems to lack a solid base of ethnographic evidence.[14] This is not surprising, for as we shall see the available data are scarce.

[14] Even though it is now routinely cited across disciplines: in *Nature* Rockenbach and Milinski (2006: 719) for example mention "direct punishment of defectors" as "a universal feature in all human societies", while in *Biology and*

Before we look at the data more carefully, however, it is worth asking what kind of evidence would support the Strong Reciprocity account of punishment and cooperation. Notice that all the quotes reported above 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 and Fehr 2004: 56, emphasis in the original)

More precisely, there are two kinds of cost that are relevant for our purposes. Let us call b_i the benefit enjoyed by an individual i from consuming a public good produced by her group. The *absolute* cost, c_i, is the fee paid by that individual (in material terms) to punish a free rider. The *relative* cost of punishment, in contrast, 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 $j \neq i$). Absolute and relative costs must be kept separate because they raise different problems for different theoretical perspectives. When sanctions are costly in absolute terms ($c_i > b_i$), punishment cannot be explained using models based on self-interested motivation. When $b_i > c_i$ punishment is consistent with self-interest, but may be problematic from an evolutionary point of view. If $b_j > (b_i - c_i)$ in fact individual selection works against the punisher: the relative cost of punishment is positive and therefore non-punishers are advantaged in fitness terms.[15] But it is also possible that $c_i > b_i$ for each individual and yet $b_j = (b_i - c_i)$, for example because the costs are spread in such a way that everybody carries an equal share of the overall burden (the absolute cost is positive but the relative cost is nil, in other words). In such a case, a motivation that overcomes the self-serving bias would not be selected against within the group.

Philosophy Sripada (2005: 782) writes that "moral norms are universally supported by punishment directed at those that violate norms".

[15] 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, for in this case the force of group selection can compensate for the adverse effect of individual selection. Weak altruism is also the basis of Sober and Wilson (1998) widely cited account of the evolution of moral norms. I'm indebted to Alejandro Rosas for this point.

9. Sex and death

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 direct 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.

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. 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.

Christopher Boehm (1999) has systematically surveyed and classified the ethnography on punishment and norm-violation.[16] Sanctions are ordered by Boehm on a scale that goes from ridicule, gossip, 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

[16] Boehm's work is also the main source of empirical evidence for Sober and Wilson (1998), which in turn is 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 the core message seems to have been lost.

cooperation (Knauft 1991), and the punishment of adulterers by jealous husbands account for a large share of murders.[17]

From an evolutionary point of view, adultery can be plausibly modelled as a Prisoner's Dilemma game with fitness payoffs,[18] and jealousy as an adaptive solution. Jealousy is a strong emotion that triggers aggressive behaviour, by-passing complex calculations of cost and benefit that may 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 wife in the future, then revenge triggered by jealousy would be advantageous from an evolutionary point of view. Punishment would not be expensive in the long run, for the punisher would recoup the costs – for example by gaining a reputation of "fierceness" that will promote access to sexual partners in the future.

The evidence unfortunately is mostly qualitative, and only suggestive. Cultures of fierceness seem more common among horticolturalists like the Yanomamo 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, to be fair.

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: 188) notes for example that dyadic club fights have the tendency to quickly escalate, and unless the elders are capable of restraining 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

[17] Among the Yanomamo, for example, "most of the club fights result from arguments over women" (Chagnon 1968: 187); "male jealousy accounts for virtually all murders among the Hadza" (Marlowe 2010: 192); and "adultery was the most common single factor" causing fights among the !Kung San (Lee 1979: 377).

[18] Using the theory of parental investment (Trivers 1972), one may argue that in a monogamous society with low rate of adultery each male has a high incentive to invest in rearing his wife's children, under the assumption that they are his own offspring. In such a society, however, male adulterers will easily spread their genes because their descendants benefit from good parental care from their non-genetic fathers. Once adultery has become common in the population, the probability that one's children are illegitimate will be high enough to deter high parental investment – and all society will be worse off in fitness terms.

aggregation. Lee (1979: 397) similarly claims that "the fear of violence … is a prominent feature of !Kung life", and these 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, instead of promoting 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. In the case of economic, rather than sexual free riding, punishment is even less common: in her study of "costly" punishment among the Ju/'hoansi, Wiessner (2005: 134) notices 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 he 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 overwhelming majority of experiments free riders cannot revenge the moralistic sanctions they have received.[19] 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).[20]

[19] Recall the empty cell in table 4: in most punishment experiments it is not possible to respond P to P.

[20] It is also interesting that in many experiments there is a non-trivial amount of anti-social sanctioning – that is, punishment aimed at *cooperators* (Cinyabuguma et al 2006, Herrmann et al 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. [21]

10. Low-cost 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 cost 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: 82-84; Ross 2006: 65-67). Going down Boehm's 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: while 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 while inflicting material punishment is likely to infringe individual rights that regulate the life of a group (e.g. property rights), symbolic punishment does not.[22]

[21] And even then, it is used only in exceptional cases, if all other forms of regulation are ineffective. Boehm suggests that overt sanctioning is so rare to be almost invisible to the external observer, which may explain why for a long time tribesmen have appeared spontaneously peaceful and non-violent to the Western eye. Acknowledgment of the rarity of material punishment is particularly significant in the writings of anthropologists like Boehm and Knauft who intend to dispel the myth of the peaceful savage.

[22] It is possible that two separate psychological mechanisms govern material and symbolic punishment: one triggered by unexpectedly harmful behaviour, eliciting aggressive punishment via anger; another one triggered by social norm violations, eliciting mockery and exclusion via "colder" negative emotions such as indignation, contempt, and disgust (Dubreuil 2010).

Moreover, 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 1998). It is a *collective* endeavour – an important point to which we 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 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 not recycling), whereas serious issues are always dealt with by "behind-your-back" gossip (Henrich and Henrich 2007: 147-150). Because of its potentially destructive effect on reputations, 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.[23]

Polly Wiessner (2005) has made a systematic attempt to find evidence of costly punishment in the field, using an extensive body of 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: 132). All in all,

[23] Rather mysteriously, Henrich and Henrich (2007) devote a lot of space in their book to Strong Reciprocity and costly punishment theory, even though their field data do not contain any evidence in their favour. All examples of reciprocity among the Chaldeans fall either in the category of repeated long-run cooperation of the Weak Reciprocity type (Henrich and Henrich have a fascinating account of how Chaldeans transform finitely repeated anonymous games into indefinitely repeated games with reputation, for example); or in the category of cooperation enforced by cheap punishment mechanisms.

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), for example, leaving one's group entails leaving one's garden, and being dependent on the hosts for food for several months.[24] 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: 248-249)

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.

[24] The guests usually pay a "rent" in terms of women.

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. In a 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 reduced the probability that an animal would be captured by the group at all.

In 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: 97-101) Cephu immediately became the victim of moralistic aggression by the group as a whole. 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.[25] 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, in 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's 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.

[25] Several ethnographers have noticed that adult males tend to refrain from overt criticism, and leave gossiping and ridiculing to women and youngsters. This division of roles is probably a form of control of male aggressiveness, and reduces the probability that verbal criticism may degenerate into violent confrontation (see also section 8 above).

> 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: 100-101)

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

The second interesting point is that the free rider is punished by taking away his illicit gain. But, *pace* Strong Reciprocity theory, no wealth is destroyed, because the other families consume what Cephu has caught. And even Cephu's punishment turns out to be not so harsh after all: once peace has been restored, one member of the main group takes some food to Cephu's hut to feed him and his family. At that point all animosity seems to be gone, and Cephu participates in the feast with the rest of the group (Turnbull 1961: 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 a handful of other episodes of moralistic aggression and ostracism reported in the anthropological literature.[26] It shows that even cases that seem favourable (because, for example, 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. [27]

[26] See e.g. Briggs (1970) and Boehm (1999, Chapter 3).

[27] Theoretical work on coalitional punishment is just beginning to emerge. Boyd et al (2010) propose a model in which the cost of punishment is inversely proportional to the number of punishers, and players can condition their decision on

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 up by material punishment work best (cf. Masclet et al 2003, Noussair and Tucker 2005, Janssen et al 2010). 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, Xiao and Houser 2005, Rockenbach and Milinski 2008, Ule et al 2009).

Most of these experiments, however, still ignore the problem of feuds and the anti-social effect of counter-punishment. There are to date only three experimental studies that combine alternative ways of incentivizing cooperation – including costly punishment – with the threat of counter-punishment. Nikiforakis and Engelmann (2008) 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. Moreover, in the aggregate costly punishment turns out to be no more efficient (and it is often less efficient) than non-costly mechanisms.[28] Although punishment pushes the rate of cooperation up, the material advantage it provides is offset by its material cost (which results in a net destruction of resources). Janssen et al (2010) similarly report a strong positive effect on revenue when communication is allowed, and when communication is matched with punishment. Costly punishment alone is not an efficient solution to social dilemmas in the laboratory. Clearly this is deeply problematic, given the Strong Reciprocity theorists' emphasis on group selection.

---

the size of the coalition. They show that under a plausible range of parameters cooperation and punishment can co-evolve.

[28] 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 and Riedl 2008, Nikiforakis and Normann 2008, Ohtsuki et al 2009). Notice that even low-cost punishment does not necessarily improve aggregate payoffs – in fact it often reduces them.

12. How common pool institutions sustain cooperation

I have described ostracism in some detail because the economists who are unfamiliar with the ethnographic 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. [29] 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 a way that has little to do with costly punishment. Rather, they tackle the problem by removing the obstacles that prevent *non*-costly mechanisms from functioning.

The emergence of institutions that regulate cooperation has been studied in depth, and 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 create "artificially" those conditions that, according to folk-theorem 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 North East 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 extensively by historians.[30] Using a data-base of over two hundred villages, Casari (2007) has shown that the charters had a

[29] Another puzzling aspect of Strong Reciprocity theorists' recent passion for anthropology is that the members of small societies on average display *less* cooperative behaviour in social dilemma games than members of large societies, and (with important variations) there seems to be a positive correlation between cooperativeness and exposure to markets (Henrich et al 2004). The key to the riddle of cooperation then is probably to be found in large societies and institutions supporting impersonal market relations, although the extent to which costly punishment plays a role in such societies is very much an open question.

[30] See e.g. McCloskey (1972).

common structure and were aimed at removing precisely the obstacles that impeded 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 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: 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.[31] 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

[31] It is also possible that insiders' transgressions were excused in particularly unlucky circumstances.

did adopt one at all.[32] The *Carte* are absolutely typical from this respect: Elinor Ostrom (1990), fresh winner of the 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: 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.

Since Ostrom's work is often cited by Strong Reciprocity theorists in support of their theses, 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. While Ostrom (1990) emphasises the cost of *setting up* common pool institutions, Strong Reciprocity theorists focus on the on-going cost of *inflicting* punishment. These two problems are quite different and should be kept distinct.[33]

Institutions such as the Trentino charters are in many ways more similar to national states in the way in which they administer sanctioning, than to the uncoordinated mechanisms of punishment experiments. Once the institutions are in place, the cost of running them (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. These advantages are what make these institutions robust and resilient across time. In costly punishment models of cooperation in contrast sanctions are a never-ending burden that must be voluntarily

[32] 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: 209-213).

[33] This is not the place to examine in detail how the problem of setting up common-pool institutions is solved in real-world cases, but Ostrom (1990) offers several illustrations. One common mechanism is that common-pool institutions "piggy-back" on previously existing institutions that were created for rather different purposes.

carried by individuals.[34] This is yet another feature that makes Strong Reciprocity fragile, on top of those discussed in previous sections.

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 can both 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 detail.

13. Models and policies

Having presented the bulk of the argument, I shall now turn to an obvious objection that can be raised against it. Lacking precise quantitative data, throughout the paper I have spoken rather liberally of "cheap", "low-cost", or "no-cost" punishment, 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 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 b especially – one should not rush to conclusions as soon as a small positive c is detected. Nevertheless, I want to argue that even small but positive *net* costs ($b - c$) 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-

[34] Yamagishi (1986) and Gurerk et al (2006) are exceptions in the experimental literature. For a theoretical analysis of "coordinated" vs. uncoordinated" punishment see Boyd et al (2010).

> 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 and Gintis 2002: 127-8)

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.[35]

How do Weak and Strong Reciprocity fare from this respect? Costly punishment experiments are often accompanied – as in the above quotation – by suggestions that self-interest, long-term horizon, and information do not matter, or matter less than traditionally assumed by economists and biologists. But this suggestion is probably 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 vs. zero-cost punishment loses much of its importance. Perhaps gossip, ostracism, and verbal reproaches *are* a bit costly, and genetic/cultural 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 you can. 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.

[35] The philosophy of science literature is very poor of insights on this important aspect of applied science; Alexandrova (2008) is a rare attempt to understand the role of models in institutional design.

14. Concluding remarks: reciprocity without costly punishment

In this paper I have argued that:

(i) two interpretations of costly punishment experiments – narrow and wide – are usually conflated by Strong Reciprocity theorists;
(ii) 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;
(iii) 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;
(iv) on the contrary, there is a lot of evidence that revenge is a major cause of dissolution of social ties;
(v) 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;
(vi) the robust common pools institutions studied by historians and institutional economists foster cooperation 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 discussed above 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.[36] 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, but its narrow focus on artificial environments in which uncoordinated costly punishment has a beneficial effect on sociality.[37]

[36] This would be a topic for a different paper; for critical perspectives on social preference theories see e.g. Binmore (2005), Bicchieri (2006), Smith (2008), Woodward (2009).

[37] Rosas (2008) argues that the same proximate (psychological) mechanisms govern both Weak and Strong negative Reciprocity. If he is right, then punishment experiments are still useful as measurement devices, even though they do not shed much light on the institutional mechanisms through which cooperation is sustained in field settings. For a different view on the psychology of reciprocity, see Dubreuil (2010).

I should also clarify that the lack of support for the costly punishment account of cooperation is not to be celebrated in my view. 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 decentralized 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 a bad idea at the level of institutional design.

Third, it is 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 gear and seek the test of historical and field data. This step was taken a long time ago in the investigation of closely 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 paper 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 much brighter at first sight. Robust support comes from surveys (Andreoni et al 1998, Fong 2001), laboratory (e.g. Berg et al 1995, Fehr et al 1993, Fischbacher et al 2001, Burlando and Guala 2005), and field experiments (Frey and Meier 2004, Shang and Croson 2005).[38]

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

[38] Notice that these are field experiments proper, not "experiments in the field" as those mentioned in earlier sections.

opportunities are processed differently from directly incurred costs (e.g. Kahneman et al 1991, Borges and Knetsch 1997). 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. Yacubian et al 2006, Tom et al 2007).

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 invested 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. 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, decentralised, uncoordinated policing are not backed up by the empirical evidence collected so far.

References

Alexandrova, A. (2008) "Making Models Count", *Philosophy of Science* 75: 383-404.

Andreoni, J., Erard, B. and Feinstein, J. (1998) "Tax Compliance", *Journal of Economic Literature* 36: 818-860.

Axelrod, R. (1984) *The Evolution of Cooperation*. London: Penguin.

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

Bardsley, N., Cubitt, R., Loomes, G., Moffatt, P., Starmer, C., and Sugden, R. (2009) *Experimental Economics: Rethinking the Rules*. Princeton: Princeton University Press.

Bendor, J. and Swistak, P. (1995) "Types of Evolutionary Stability and the Problem of Cooperation", *Proceedings of the National Academy of Science* 92: 3596-3600.

Bendor, J. and Swistak, P. (1997) "The Evolutionary Stability of Cooperation", *American Political Science Review* 91: 290-307.

Berg, J., Dickhaut, J. and McCabe, K. (1995) "Trust, Reciprocity, and Social History", *Games and Economic Behaviour* 10, 122-142.

Bergstrom, T.C. (2002) "Evolution of Social Behavior: Individual and Group Selection", *Journal of Economic Perspectives* 16: 67-88.

Bicchieri, C. (2006) *The Grammar of Society*. New York: Cambridge University Press.

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

Binmore, K. (1999) "Why Experiment in Economics?" *Economic Journal* 109: F16–24.

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

Binmore, K. (2006) "Why Do People Cooperate?" *Politics, Philosophy and Economics* 5: 81-96.

Boehm, C. (1999) *Hierarchy in the Forest: The Evolution of Egalitarian Behavior*. Cambridge, Mass.: Harvard University Press.

Borges, B.F.J. and Knetsch, J.L. (1997) "Valuation of Gains and Losses, Fairness, and Negotiation Outcomes", *International Journal of Social Economics* 24: 265-281.

Bowles, S. and Gintis, H. (2002) "Homo Reciprocans", *Nature* 415: 125.128.

Bowles, S. and Gintis, H. (2003) "The Evolution of Strong Reciprocity: Cooperation in Heterogeneous Populations", *Theoretical Population Biology* 65: 17-28.

Boyd, R. and Richerson, P. (1990) "Group Selection among Alternative Evolutionarily Stable Strategies", *Journal of Theoretical Biology* 145: 331-342.

Boyd, R. and Richerson, P. (1992) "Punishment Allows the Evolution of Cooperation (Or Anything Else) in Sizable Groups", *Ethology and Sociobiology* 13: 171-195.

Boyd, R., Bowles, S. and Gintis, H. (2010) "Coordinated Punishment of Defectors Sustains Cooperation and Can Proliferate when Rare", *Science* 328: 617-620.

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

Briggs, J.L. (1970) *Never in Anger: Portrait of an Eskimo Family*. Cambridge, Mass.: Harvard University Press.

Burlando, R.M. and Guala, F. (2005) "Heterogeneous Agents in Public Goods Experiments", *Experimental Economics* 8: 35-54.

Camerer, C.F. (2003) *Behavioral Game Theory*. Princeton: Princeton University Press.

Camerer, C.F. and Fehr, E. (2004) "Measuring Social Norms and Preferences Using Experimental Games: A Guide for Social Scientists", in Henrich et al (eds.) *Foundations of Human Sociality*. New York: Oxford University Press.

Casari, M. (2007) "Emergence of Endogenous Legal Institutions: Property Rights and Community Governance in the Italian Alps", *Journal of Economic History* 67: 191-226.

Chagnon, N.A. (1968) *Yanomamö: The Fierce People*. New York: Holt, Rinehart and Winston, 6th edition 1992.

Chagnon, N.A. (1988) "Life Histories, Blood Revenge, and Warfare in a Tribal Population", *Science* 239: 985-992.

Cinyabuguma, M., Page, T. and Putterman, L. (2004) "Cooperation under the threat of expulsion in a public goods experiment", *Journal of Public Economics* 89: 1421-1435.

Cinyabuguma, M., Page, T. and Putterman, L. (2006) "Can Second-Order Punishment Deter Perverse Punishment?" *Experimental Economics* 9: 265-279.

Cooper, D.J. and Dutcher, E.G. (2009) "The Dynamics of Responder Behavior in Ultimatum Games: A Meta-study", Working Paper, Florida State University.

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

Denant-Boemont, L., Masclet, D. and Noussair, C. (2007) "Punishment, Counterpunishment and Sanction Enforcement in a Social Dilemma Experiment", *Economic Theory* 33: 145-167.

de Quervain D.J.F., Fischbacher U., Treyer V., Schellhammer M., Schnyder U., Buck A. and Fehr E. (2004) "The Neural Basis of Altruistic Punishment", *Science* 305, 1254-1258.

Dreber, A., Rand, D.G., Fudenberg, D. and Nowak, M.A. (2008) "Winners Don't Punish", *Nature* 452: 348-351.

Dubreuil, B. (2010) "Punitive Emotions and Norm Violations", *Philosophical Explorations* 13: 35-50.

Dunbar, R. (1998) *Grooming, Gossip, and the Evolution of Language*. Cambridge, Mass.: Harvard University Press.

Egas, M. and Riedl, A. (2008) "The Economics of Altruistic Punishment and the Maintenance of Cooperation", *Proceedings of the Royal Society B* 275: 871-878.

Falk, A. and Fischbacher, U. (2005) "Modeling Strong Reciprocity", in Gintis, H., Boyd, R., Bowles, S. and Fehr, E. (eds.) *Moral Sentiments and Material Interests*. Cambridge, Mass.: MIT Press.

Falk, A., Fehr, E. and Fischbacher, U. (2003) "On the Nature of Fair Behavior", *Economic Inquiry* 41: 20-26.

Fehr, E. and Fischbacher, U. (2002) "Why Social Preferences Matter – The Impact of Non-Selfish Motives on Competition, Cooperation and Incentives", *Economic Journal* 112: C1-C33.

Fehr, E. and Fischbacher, U. (2004) "Third-Party Sanctions and Social Norms", *Evolution and Human Behavior* 25: 63-87.

Fehr, E. and Fischbacher, U. (2005) "The Economics of Strong Reciprocity", in Gintis, H., Boyd, R., Bowles, S. and Fehr, E. (eds.) *Moral Sentiments and Material Interests*. Cambridge, Mass.: MIT Press.

Fehr, E. and Gachter, S. (2000) "Cooperation and Punishment in Public Goods Experiments", *American Economic Review* 90: 980-994.

Fehr, E. and Gachter, S. (2002) "Altruistic Punishment in Humans", *Nature* 415: 137-140.

Fehr, E. and Schmidt, K. (2006) "The Economics of Fairness, Reciprocity and Altruism - Experimental Evidence and New Theories", in *Handbook of the Economics of Giving, Reciprocity and Altruism*, Amsterdam: Elsevier.

Fehr, E., Kirchsteiger, G. and Riedl, A. (1993) "Does Fairness Prevent Market Clearing? An Experimental Investigation". *Quarterly Journal of Economics* 108: 437-460.

Fischbacher, U., Gachter, S., and Fehr, E. (2001) "Are People Conditionally Cooperative? Evidence from a Public Goods Experiment." *Economics Letters* 71: 397-404.

Fong, C. (2001) "Social Preferences, Self-interest, and the Demand for Redistribution", *Journal of Public Economics* 82: 225-246.

Frank, R.H. (1988) *Passions within Reason: the Strategic Role of Emotions*. New York: Norton.

Frey, B. and Meier, S. (2004) "Social Comparison and Pro-social Behavior: Testing 'Conditional Cooperation' in a Field Experiment", *American Economic Review* 94: 1717-1722.

Fudenberg, D., Levin, D.K. and Maskin, E. (1994) "The Folk Theorem with Imperfect Public Information", *Econometrica* 62: 997-1039.

Fudenberg, D. and Maskin, E. (1986) "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information", *Econometrica* 54: 533-54.

Gintis, H. (2000) "Strong Reciprocity and Human Sociality", *Journal of Theoretical Biology* 206: 169-179.

Gintis, H. (2006) "Behavioral Ethics Meets Natural Justice", *Politics, Philosophy and Economics* 5: 5-32.

Gintis, H. (2009) *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton: Princeton University Press.

Gintis, H., Boyd, R., Bowles, S. and Fehr, E. (2003) "Explaining Altruistic Behavior in Humans", *Evolution and Human Behavior* 24: 153-172.

Gintis, H., Boyd, R., Bowles, S. and Fehr, E. (eds. 2005) *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*. Cambridge, Mass.: MIT Press.

Gouldner, A.W. (1960) "The Norm of Reciprocity: A Preliminary Statement", *American Sociological Review* 25: 161-178.

Guala, F. (2005) *The Methodology of Experimental Economics*. New York: Cambridge University Press.

Guala, F. (2008) "Paradigmatic Experiments: The Ultimatum Game from Testing to Measurement Device", *Philosophy of Science* 75: 658-669.

Gürerk, O., Irlenbusch, B. and Rockenbach, B. (2006) "The Competitive Advantage of Sanctioning Institutions", *Science* 312: 108-111.

Gurven, M. (2004) "To Give and to Give Not: The Behavioral Ecology of Human Food Transfers", *Behavioral and Brain Sciences* 27: 543-583.

Güth, W., Schmittberger, R., and Schwarz, B. (1982), "An Experimental Analysis of Ultimatum Bargaining", *Journal of Economic Behavior and Organization* 3: 367-388.

Hacking, I. (1988) "The Participant Irrealist at Large in the Laboratory", *British Journal for the Philosophy of Science* 39: 277-294.

Hagen, E. H. and Hammerstein, P. (2006) "Game Theory and Human Evolution: A Critique of Some Recent Interpretations of Experimental Games", *Theoretical Population Biology* 69: 339-348.

Harrison, G. W. and List, J. A. (2004) "Field Experiments," *Journal of Economic Literature* 42: 1009-45.

Henrich, J. and Boyd, R. (2001) "Why People Punish Defectors: Weak Conformist Transmission can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas", *Journal of Theoretical Biology* 208: 79-89.

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

Henrich, J. and Henrich, N. (2007) *Why Humans Cooperate: A Cultural and Evolutionary Explanation*. New York: Oxford University Press.

Herrmann, B., Thoni, C. and Gachter, S. (2008) "Antisocial Punishment across Societies", *Science* 319: 1362-1367.

Hirshleifer, J. (1987) "On the Emotions as Guarantors of Threats and Promises", in Dupré, J. (ed.) *The Latest on the Best.* Cambridge, Mass.: Harvard University Press.

Janssen, M.A., Holahan, R., Lee, A. and Ostrom, E. (2010) "Lab Experiments for the Study of Social-Ecological Systems", *Science* 328: 613-617.

Kahneman, D., Knetsch, J.L. and Thaler, R.H. (1991) "Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias", *Journal of Economic Perspectives* 5: 193-206.

Knauft, B.M. (1987) "Reconsidering Violence in Simple Human Societies: Homicide among the Gebusi of New Guinea", *Current Anthropology* 28: 457-500.

Knauft, B.M. (1991) "Violence and Sociality in Human Evolution", *Current Anthropology* 32: 391-428.

Lee, R.B. (1979) *The !Kung San: Men, Women, and Work in a Foraging Society*. Cambridge: Cambridge University Press.

Mahdi, N.Q. (1986) "Pukhtunwali: Ostracism and Honor among the Pathan Hill Tribes", *Ethology and Sociobiology* 7: 295-304.

Marlowe, F.W. (2010) *The Hadza: Hunter-Gatherers of Tanzania.* Berkeley: University of California Press.

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: 587-592.

Marshall, L. (1961) "Sharing, Talking, Giving: Relief of Social Tensions among the !Kung Bushmen", *Africa* 31: 231-249.

Masclet, D., Noussair, C., Tucker, S. and Villeval, M.C. (2003) "Monetary and Nonmonetary Punishment in the Voluntary Contributions Mechanism", *American Economic Review* 93: 366-380.

Mauss, M. (1954) *The Gift: Forms and Functions of Exchange in Archaic Societies*. London: Cohen & West.

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.

Nikiforakis, N. (2008) "Punishment and Counter-Punishment in Public Good Games: Can We Really Govern Ourselves?" *Journal of Public Economics* 92: 91-112.

Nikiforakis, N. and Engelmann, D. (2008) "Feuds in the Laboratory? A Social Dilemma Experiment", Research Paper 1058, Department of Economics, University of Melbourne.

Nikiforakis, N. and Normann, H.T. (2008) "A Comparative Statics Analysis of Punishment in Public-Good Experiments", *Experimental Economics* 11: 358-369.

Noussair, C. and Tucker, S. (2005) "Combining Monetary and Social Sanctions to Promote Cooperation", *Economic Inquiry* 43: 649-660.

Ostrom, E. (1990) *Governing the Commons: The Evolution of Institutions for Collective Action*. New York: Cambridge University Press.

Ostrom, E. (2000) "Collective Action and the Evolution of Social Norms", *Journal of Economic Perspectives* 14: 137-158.

Ostrom, E., Walker, J. and Gardner, R. (1992) "Covenants with and without a Sword: Self-Governance Is Possible", *American Political Science Review* 86: 404-417.

Ohtsuki, H., Iwasa, Y., Nowak, M.A. (2009) "Indirect Reciprocity Provides Only a Narrow Margin of Efficiency for Costly Punishment". *Nature* 457: 79–82.

Page, T. Putterman, L. and Unel, B. (2005) "Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry, and Efficiency", *Economic Journal* 115: 1032-1053.

Richerson, P.J. and Boyd, R. (2005) *Not By Genes Alone: How Culture Transformed Human Evolution*. Chicago: University of Chicago Press.

Rockenbach, B. and Milinski, M. (2006) "The Efficient Interaction of Indirect Reciprocity and Costly Punishment", *Nature* 444: 718-723.

Rosas, A. (2008) "The Return of Reciprocity: A Psychological Approach to the Evolution of Cooperation", *Biology and Philosophy* 24: 555-566.

Ross, D. (2006) "Evolutionary Game Theory and the Normative Theory of Institutional Design: Binmore and Behavioral Economics", *Politics, Philosophy, and Economics* 5: 51-79.

Roth, A., Prasnikar, V., Okuno-Fujiwara, M. and Zamir, S. (1991), "Bargaining and Market Behavior in Jerusalem, Lubljana, Pittsburgh and Tokyo: An Experimental Study", *American Economic Review* 81: 1068-1095.

Sahlins, M. (1974) *Stone Age Economics*. London: Routledge.

Sanfey, A.G., Rilling, J.K., Aaronson, J.A., Nystrom, L.E., and Cohen, J.D. (2003) "The Neural Basis of Economic Decision-Making in the Ultimatum Game." *Science*, 300: 1755–58.

Shang, J. and Croson, R. (2005) "The Impact of Social Comparison on Contribution Decisions: Mailing Experiments, Surveys and Laboratory Tests". *Working Paper, University of Pennsylvania*.

Singer T. and Fehr E. (2005) "The Neuroeconomics of Mind Reading and Empathy", *American Economic Review* 95: 340-345.

Smith, V.L. (2008) *Rationality in Economics: Constructivist and Ecological Forms*. New York: Cambridge University Press.

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

Spitzer M., Fischbacher U., Herrnberger B., Gron G. and Fehr E. (2007) "The Neural Signature of Norm Compliance", *Neuron* 56: 185-196.

Sripada, C.S. (2005) "Punishment and the Strategic Structure of Human Systems", *Biology and Philosophy* 20: 767-789.

Starmer, C. (1999) "Experiments in Economics … (Should We Trust the Dismal Scientists in White Coats?)" *Journal of Economic Methodology* 6: 1-30.

Steel, D. (2007) *Across the Boundaries: Extrapolation in Biology and in the Social Sciences*. New York: Oxford University Press.

Tom, S.M., Fox, C.R., Trepel, C., and Poldrack, R.A. (2007) "The Neural Basis of Loss Aversion in Decision-Making Under Risk", *Science* 315: 515 – 518.

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

Trivers, R.L. (1972) "Parental Investment and Sexual Selection", in B.G. Campbell (ed.) *Sexual Selection and the Descent of Man*. Chicago: Aldine Pub.

Trivers, R.L. (2004) "Behavioural Evolution: Mutual Benefits at All Levels of Life", *Science* 304: 964-965.

Turnbull, C. (1961) *The Forest People*. London: Jonathan Cape.

Ule, A., Schram, A., Riedl, A., and Cason, T.N. (2009) "Indirect Punishment and Generosity toward Strangers", *Science* 326: 1701-1704.

Wiessner, P. (2005) "Norm Enforcement among the Ju/'hoansi Bushmen", *Human Nature* 16: 115-145.

Wilson, D.S. (1979) "Structured Demes and Trait-Group Variation", *American Naturalist* 113: 606-610.

Wilson, D.S. and Sober, E. (1994) "Reintroducing Group Selection to the Human Behavioral Sciences", *Behavioral and Brain Sciences* 17: 585-654.

Woodward, J. (2009) "Experimental Investigations of Social Preferences", in H. Kincaid and D. Ross (eds.) *The Oxford Handbook of Philosophy of Economics*. New York: Oxford University Press, pp. 189-222.

Xiao, E. and Houser, D. (2005) "Emotion Expression in Human Punishment Behavior", *Proceedings of the National Academy of Science* 102: 7398-7401.

Yacubian, J., Gläscher, J., Schroeder, K., Sommer, T., Braus, D,F. and Büchel, C. (2006) "Dissociable Systems for Gain- and Loss-Related Value Predictions and Errors of Prediction in the Human Brain", *Journal of Neuroscience* 26: 9530-9537.

Yamagishi, T. (1986) "The Provision of a Sanctioning System as a Public Good", *Journal of Personality and Social Psychology* 51: 110-116.