Author's Response

Strong Reciprocity Is Real, but There Is No Evidence that Uncoordinated Costly Punishment Sustains Cooperation in the Wild

Francesco Guala
Department of Economics, University of Milan, via Conservatorio 7, 20122 Milan, Italy.
Francesco.Guala@unimi.it

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

R1. Introduction

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

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

1) Strong negative Reciprocity is a real proximate cause of human behaviour, and may *indirectly* promote cooperation.
2) Punitive motives in particular sustain institutions that administer sanctions, satisfying the moralistic preferences of the cooperative members of society.

3) To be viable, however, these institutions typically reduce the costs of sanctions and in particular prevent the eruption of feuds triggered by uncoordinated punishment.
4) Uncoordinated costly punishment thus is unlikely to be a *direct* mechanism sustaining cooperation in the wild. Successful societies either find ways to administer coordinated punishment at low cost, or abstain from punishing free riders as documented by anthropologists.

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

R2. Is Strong Reciprocity a straw man?

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

In my target article I define Strong Reciprocity as the propensity to reply "nice to nice" and "nasty to nasty" actions, even if this entails a net cost for the individual agent. The Strong Reciprocity *programme* as I understand it aims at explaining various aspects of human sociality using models that incorporate Strong positive or Strong negative Reciprocity motives (sometimes called "social preferences") among their premises. The *costly punishment hypothesis* which is the main focus of

my target article is an important element of Strong Reciprocity theory, i.e. the idea that uncoordinated costly sanctions supported by Strong negative Reciprocity efficiently discipline free riders and protect positive reciprocators in social dilemma games.

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

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

R3. Are Weak and Strong Reciprocity mutually exclusive?

As evidence that I mischaracterize their programme, **Henrich & Chudek** cite several papers in which Strong Reciprocity theorists appeal to Weak Reciprocity mechanisms (such as kinship, or reputation) to explain human cooperation. But nowhere in my target article I say that Weak and Strong Reciprocity are incompatible, or I attribute this claim to Strong Reciprocity theorists. The controversial issue is whether Strong Reciprocity has a significant effect on cooperation, over and beyond the effect of Weak Reciprocity.

There is plenty of evidence in support of Weak Reciprocity, as several commentators highlight. Cultural factors (e.g. norms) are favoured by many (e.g. **Boehm**, **Henrich & Chudek**, **Read, von**

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

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

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

R4. Can field data solve the riddle of cooperation?

An important methodological thesis of my article is that laboratory data alone cannot solve the riddle of cooperation. It is easy to take this as a general anti-experimental argument, which is why perhaps some commentators felt the need to stress that field data alone are not enough either. I agree: my paper was never meant to be an anti-experimental treatise, as **Casari** and **Nikiforakis** seem to have interpreted. I am very fond of laboratory experiments, a methodology that has greatly enriched the toolbox of social science and human biology (Guala 2005). My suggestion is, more modestly, that in this particular juncture we will gain more by combining what we have learned in the laboratory with the message of field studies. This in turn will lead to more (and better) experiments, which together with new field data will drive progress in this difficult but fascinating field of research. **Nikiforakis** and **Casari** already use this eclectic approach in their work. **Bereby-Meyer** and **Pisor & Fessler** provide further examples of how it is possible to bring more realism in experiments, or how to incorporate in experimental designs some features that are typical of the real-world circumstances in which cooperation takes place. My trivial point is that in order to run more realistic experiments, we need to know more about cooperation in the wild.

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

R5. Does ethnographic evidence support the costly punishment hypothesis?

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

Of the old literature, **Gintis & Fehr** keep citing the work of Boehm and Wiessner. In my article I argue that the evidence reported in these studies does not support the costly punishment story. (One

of the articles by Wiessner (2009), by the way, says explicitly so.) The commentaries published in this issue of *BBS* support my interpretation: **Wiessner** agrees that "experimental and ethnographic evidence do not concur", and **Boehm** similarly claims that "the costs do not fit necessarily with assumptions made in models that consider punishment to be altruistic". Other anthropologists (e.g. **Baumard**, **von Rueden & Guerven**, and **Read)** argue that there are plausible alternative readings of the evidence. (**Read** in fact says that I do not go far enough in my discussion of the "disjunction between experimental and real conditions".)

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

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

What we need are especially customized new measurements, where all the obvious confounds have been estimated and tested using rigorous statistical techniques.

R6. Have I overlooked some field data in favour of Strong Reciprocity?

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

Casari says that costly punishment is still practised in Trentino, the region at the centre of his research on the *Carte di regola*. He mentions damages to young grapevines carried out at night, but not enough detail is provided to figure out what these stealth expeditions are really about. Are they meant to enforce cooperation in the pursuit of a common good? Or are they just petty jealousies among neighbours? Are these *individual* initiatives (perhaps part of ongoing feuds) or *coordinated* actions backed up by the whole community? These questions are crucial, because as I have said the existence of punishment is not at issue here nor, similarly, in the ethnographic literature. The issue is whether such punishment is costly to individual punishers, and whether it sustains or disrupts social cooperation.

R7. Is evidence for Strong Reciprocity hard to find because costly punishment is rare?

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

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

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

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

The emphasis on error is quite important, as highlighted in **Dreber & Rand**'s commentary. What looks like an equilibrium when error is not permitted, may turn out to be unstable in a stochastic environment. Uncertainty is likely to play an important role in real-world environments – recall the one of the complaints of Strong Reciprocity theorists is that the almost perfect monitoring required by folk theorems is unrealistic. **Bereby-Meyer** notices that the introduction of uncertainty in ultimatum games reduces the rate of rejections significantly (see also **van Lange, Balliet & IJzerman** for related comments). This might explain why punishment is observed only rarely in the field, but it is a rather different type of explanation from **Gächter**'s: if people give others the benefit

of the doubt, free riding becomes more profitable and sanctions *less* effective. In section 13 of the target article I explain how successful institutions help solve this problem, by coordinating monitoring and resolving whatever uncertainties there may be (for example, on the interpretation of rules). While the existence of such institutions is almost certainly backed up by Strong Reciprocity motives, their smooth functioning relies on Weak Reciprocity mechanisms that guarantee long-term profitability, sustainability, and efficacy.

R8. Does punishment have to be uncoordinated?

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

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

Bowles, Boyd, Mathew & Richerson say that it is illegitimate to suppose that the cost of punishment ought to be large. But how large is "large" in this context? The 1/3 ratio between cost and inflicted damage that is used in many experiments is unrealistic for situations in which punishment can be retaliated by equally strong individuals. But even the 1/3 ratio generates inefficient outcomes (e.g. Egas & Riedl 2008). In such circumstances, either cooperation is bound

to collapse, or people must devise cheaper ways to enforce it. I suspect that both cases are common, but the study of successful resilient institutions suggests that if there are superior alternatives to uncoordinated costly punishment people tend to exploit them. So I agree with **Casari** that one important reason why costly punishment is not frequently observed in the field is that people find better ways to enforce cooperation.

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

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

Van den Berg, Molleman & Weissing cite costs generated by ostracism that can easily be overlooked, like the creation of predatory outcasts ("desperados") or the disruption of social relations that are crucial for a well-functioning group. While I agree with them that further research is required on these costs and their quantitative impact, I should point out that their existence is well known to ethnographers. **Boehm**, for example, describes various mechanisms observed in small

societies that have the effect of distributing the costs of sanctions over the members of the group and of alleviating some side-effects of punishment. Kinsmen are chosen to act as punishers or peacemakers; the identity of executioners is kept secret, or the group as a whole acts as killer. He also notices that such mechanisms are determined culturally and situationally, which reduces the problem of (genetic) free riding.

I do not have the expertise to comment on the importance of these cultural mechanisms, but I have no doubt that we ought to study them in more detail. One important message of my target article is that it is time to abandon anecdotal evidence and move on to quantitative analysis. The assassination of mobsters mentioned by **Runciman**, unless articulated in further detail, belongs to the realm of the anecdote. As Gambetta (1996) explains convincingly, trust and reputation (i.e. Weak Reciprocity) are crucial cogwheels in the functioning of the Sicilian Mafia. And the very strategy of costly signalling mentioned by **Runciman** can be explained using standard game-theoretic models based on Nash equilibrium, in which the costs are recouped later in the game. **Runciman** is right, I believe, to say that in every successful institution "strong reciprocity is waiting in the wings … to ensure the covenant is renewed". My purpose is not to deny that Strong Reciprocity motives exist (see also R13 below), but to point out that there is no evidence that they sustain cooperation by way of uncoordinated costly punishment in the wild.

R10. Are costs recouped via group selection?

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

This argument relies on group selection, which is itself a controversial theory in evolutionary biology. According to one prominent interpretation, group selection models are just special cases of standard models based on inclusive fitness and kin selection, that apply when certain parameters take extreme values (for example when within-group competition is very low – see West, Mouden & Gardner 2010, for a recent statement). Under this interpretation, then, **Henrich & Chudek** are right that Weak and Strong Reciprocity explanations do not differ radically. However all the

objections to a costly punishment account of field data that I present in my article identify some mechanism (like reputation, coalitional punishment, etc.) that reduces the relative costs *within* the group. If the objections are sound, the free rider problem may be negligible or non-existent, and there may be no need to recoup the costs at a higher level via group selection. This does not mean that competition between groups is not important, of course; only that it might not solve *this* particular problem. (In fact the opposite is likely to be true: group selection works more smoothly if the free rider problem within each group has already been solved using non-costly punishment mechanisms – see e.g. Sober & Wilson 1998).

In my target article I do not put much emphasis on group selection because it plays an ancillary role in this debate. Scholars in both camps agree that at *some* level the costs have to be recouped. The contentious issue is where: if punishment is costly, then group selection has a lot of work to do; if it is not, group selection may have an easier job or (perhaps) no job at all.

R11. Are experiments good predictors of field behaviour?

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

Bowles, Boyd, Mathew & Richerson, **Ferguson & Corr**, **Henrich & Chudek**, and **Johnson** highlight the positive correlations as proof that laboratory behaviour predicts (at least partially) behaviour in the field. **Civai**, **Pisor & Fessler**, **von Rueden & Gurven**, and **Wiessner** in contrast highlight the *lack* of correlation found in other studies as evidence of the importance of contextual factors. My view is that we need a systematic analysis of when, where, and why such correlations obtain, before we can say anything general about the power of experiments as predictors of non-laboratory behaviour. Focusing on successes (e.g. positive correlations) may be justified at an early

stage of research, when one is looking for surprising results, but at a later stage it must be supplanted by a quantitative assessment of successes and failures.

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

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

Having said that, let me emphasize that I never meant to claim that the results of experimental games have no external validity. On the contrary, I believe they do in a number of cases. In fact it would be surprising to find no correlation between the behaviour in and out of the lab. My external validity worry is different: uncoordinated costly punishment may be a bad solution to the problem of cooperation because in realistic environments it creates more problems than it solves. That's why societies have found alternative ways to sustain cooperation and to harness the natural impulse to

sanction free riding. The problem is not that Strong negative Reciprocity occurs in the lab only: on the contrary, because it is a real force everywhere, it has to be carefully managed, channelled, and if necessary suppressed.

R12. Should we talk about ultimate causes only?

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

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

This unification has rather unpleasant implications for Weak Reciprocity theory, though, because models based on selfish preferences and strategic reasoning are too limited to account for the variety of proximate causes of human behaviour (**Rosas**). One solution is to retreat to an "as if" interpretation of these models, and defend them as useful instruments based on unrealistic assumptions. Although there is an old instrumentalist tradition in economics, "as if" interpretations have been used too often to shield theories from criticism. In contrast, models based on false principles should be modified to build better proximate models consistent with the spirit of Weak Reciprocity. The success of folk theorem-like explanations prompts us to ask how such idealized models can nevertheless be useful as stylized explanations – a question that has puzzled many

scientists and philosophers since David Hume formulated it three centuries ago. But in search for better models one does not have to ditch the promising features of Weak Reciprocity explanations (like the emphasis on repeated play or reputation).

R13. Are pro-social motives real?

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

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

R14. Why does it matter?

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

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

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

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

Something similar might be true of Strong Reciprocity. There is a wide gap between theoretical relevance and application, and we should better acknowledge that Strong Reciprocity theory has not bridged it yet.

References

Balliet, D., Moulder, L.B. & van Lange, P.A.M. (2011) Reward, Punishment, and Cooperation: A Meta-Analysis. *Psychological Bulletin*, on-line first. http://dx.doi.org/10.1037/a0023489

Boyd, R., Gintis, H. & Bowles, S. (2010) Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science* 328:617–20. http://www.sciencemag.org/cgi/content/full/328/5978/617

Casari, M. & Luini, L. (2009) Cooperation under alternative punishment institutions: An experiment. *Journal of Economic Behavior & Organization* 71:273–82. http://dx.doi.org/10.1016/j.jebo.2009.03.022

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

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

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

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

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

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

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D. P. & Ziker, J. (2010) Market, religion, community size and the evolution of fairness and punishment. *Science 327*: 1480-1484. http://dx.doi.org/10.1126/science.1182238

Henrich, J., Heine, S.J. & Norenzayan, A. (2010) The weirdest people in the world? *Behavioral & Brain Sciences* 33: 61-83. http://dx.doi.org/10.1017/S0140525X0999152X

Mathew, S. & Boyd, R. (unpublished) Punishment sustains large-scale cooperation in pre-state warfare. Working Paper.

Meggitt, M. (1962) *Desert people: A study of the Walbiri aborigines of Central Australia*. University of Chicago Press.

Ostrom, E., Walker, J. & Gardner, R. (1992) Covenants with and without a sword: Self-governance is possible. *American Political Science Review* 86:404–17. http://www.jstor.org/stable/1964229

Putnam, R. (2001) *Bowling alone: The collapse and revival of American community.* Simon & Schuster.

Rustagi, D., Engel, S. & Kosfeld, M. (2010) Conditional cooperation and costly monitoring explain success in forest commons management. *Science* 330: 961-5. http://www.sciencemag.org/content/330/6006/961

Sober, E. & Wilson, D. S. (1998) *Unto others: The evolution and psychology of unselfish behavior.* Harvard University Press.

Strehlow, T.G.H. (1970) Geography and the totemic landscape in Central Australia: A functional study. In: *Australian aboriginal anthology: Modern studies in the social anthropology of the Australian aborigines*. ed. R. M. Berndt, University of Western Australia Press.

West, S.A., Mouden, C.E. & Gardner, A. (2010) Sixteen misconceptions about the evolution of cooperation in humans. *Evolution & Human Behavior* 32: 231-262. http://dx.doi.org/10.1016/j.evolhumbehav.2010.08.001

Wiessner, P. (2009) Experimental games and games of life among the Ju/'hoan bushmen. *Current Anthropology* 50:133–38. http://www.jstor.org/stable/20479691

Yamagishi, T. (1986) The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology* 51:110–16. http://linkinghub.elsevier.com/retrieve/pii/S0022351407601885