Why Sanction? Functional Causes of Punishment and Reward

Pat Barclay
Department of Psychology
University of Guelph
50 Stone Rd. E.
Guelph, ON, N1G 2W1, Canada
Phone: 519-824-4120 ext. 58247
Fax: 519-837-8629
barclayp@uoguelph.ca

Toko Kiyonari
School of Social Informatics
Aoyama Gakuin University
5-10-1, Fuchinobe, Chuo-ku, Sagamihara,
Kanagawa, 252-5258, Japan
Phone: +81-42-759-6114
Fax: +81-42-759-6075
kiyonari@si.aoyama.ac.jp

To appear in P. van Lange, B. Rockenbach, & T. Yamagishi (Eds.)
Social Dilemmas: New Perspectives on Reward and Punishment

Abstract

Many disciplines find it puzzling that costly cooperation exists within groups of non-kin. Cooperation can be sustained when non-cooperators are punished or when cooperators are rewarded, but these sanctions are themselves costly to provide. As such, we must ask: what forces maintain the existence of sanctioning? Why do people possess a psychology that includes punitive sentiment and a willingness to reward? Many theoretical models rely on "second-order punishment", which means that people will punish those who do not punish non-cooperators. However, our review of the evidence suggests that people do not readily do this, and do not particularly like punishers. This calls into question any theories that rely on this second-order punishment. By contrast, people will readily reward those who reward cooperators, which suggests that rewards may function as part of a system of indirect reciprocity where cooperators (and rewarders) are seen as "good" thus worthy of help. So what does sustain punitive sentiment? Punishment may function to signal qualities of the punisher that are otherwise difficult to observe, such as the punisher's trustworthiness or willingness to retaliate against personal affronts. Alternately, punishment may simply be a "Volunteer's Dilemma" where it becomes rational to "volunteer" to punish non-cooperators if no one else in the group will. Finally, we discuss how positive and negative sanctions may function differently to maintain large-scale human cooperation, depending on the proportion of cooperators in a population and the necessity for unanimous cooperation.

Keywords: cooperation, altruism, second-order punishment, indirect reciprocity, signaling

Introduction

Cooperation is a theoretical puzzle in many disciplines. Why would one organism do things that benefit others if doing so is costly? Much research has focused on the psychological mechanisms underlying such help (e.g. Andreoni, 1990; Batson et al., 1997; Cialdini, Brown, Lewis, Luce, & Neuburg, 1997; Cox, 2004). From a functional perspective, why do such cooperative sentiments persist? Why do humans possess psychological mechanisms which cause them to help others? What selective forces cause these sentiments to evolve and/or be learned? Evolutionary researchers often seek to answer these questions of function. By understanding the function(s) of cooperation, it allows us to create situations that will promote cooperation regardless of what specific sentiments trigger cooperation within each individual (Barclay, in press).

Many explanations about the functions of cooperation rely on people being able to target their cooperation towards specific individuals and away from others. For example, reciprocity involves helping those who have specifically helped you (direct reciprocity: Axelrod, 1984; Trivers, 1971) or who have been helpful in general (indirect reciprocity: Alexander, 1987; Nowak & Sigmund, 2005), and avoiding helping people who do not help. However, what happens when cooperative acts cannot be targeted towards specific individuals?

One case where cooperative acts cannot be preferentially targeted towards cooperators is with the provision of public goods. A public good is something that is costly to provide, but once provided, is available to all people regardless of how much each person contributed towards the provision (e.g. Davis & Holt, 1993; Ledyard, 1995). Because all members have equal access to the public good, those who pay the cost of providing public goods will have a lower payoff than non-contributors who do not pay this cost (a.k.a. “free-riders” or defectors). This incentive structure will often result in the underprovision of public goods, a result which has been found in many experiments (see reviews by Ledyard, 1995). A similar situation occurs with resource management because conservation of resources is a public good that everyone benefits from, and a “tragedy of the commons” can occur as individuals follow their rational incentives to overexploit common resources (Hardin, 1968). It is important to solve these twin problems of public goods provision and tragedies of the commons, especially because overuse of natural resources has been implicated in the collapse of some historical societies like the Mayans and Easter Islanders (Diamond, 2005). We will treat these problems equivalently in this chapter because of their similarity in incentive structure.

As this collection is about peer-to-peer rewards and punishment, readers will not be surprised that these have both been put forth as forces that maintain contributions to public goods. Contributing to public goods can be worthwhile if one can be rewarded sufficiently for doing so. Similarly, free-riding on others’ contributions is not worthwhile if one would face punishment for not contributing oneself. Much research has shown that people are willing to pay to sanction

others, either through rewards or through punishment, and the majority of studies show that both rewards and punishment can sustain cooperation under some circumstances (e.g. Fehr & Gächter, 2002; Gächter, Renner, & Sefton, 2008; Masclet, Noussair, Tucker, & Villeval., 2003; McCusker & Carnevale, 1995; Nikiforakis & Normann, 2008; Ostrom, Walker, & Gardner, 1992; Rand, Dreber, Ellingsen, Fudenberg, & Nowak, 2009; Rockenbach & Milinski, 2006; Sefton, Shupp, & Walker, 2007; Vyrastekova & van Soest, 2008; Yamagishi, 1986)[1]. We will let others provide a more comprehensive review of the existence and effectiveness of punishment and reward, and will instead focus on why people possess such punitive and rewarding sentiment. We will also focus on sanctions provided by individuals ("peer-to-peer sanction") instead of by institutions, as the two types of sanctions may involve different etiologies.

Goals of This Chapter

In this chapter, we will not discuss which specific psychological mechanisms (e.g. which emotions) cause cooperative actions, nor will we discuss which specific emotions cause punishment and reward. These are excellent topics and have been investigated by others (e.g. de Quervain et al., 2004; Fehr & Gächter, 2002; Hopfensitz & Reuben, 2009; Nelissen & Zeelenberg, 2009) and will likely be covered elsewhere in this volume. Instead, we ask: why does punitive sentiment exist at all? Given that it is costly to punish or reward others, why do people possess a psychology which makes them inclined to reward those who provide public goods and punish those who do not? What forces select for and maintain such sentiments? These questions of function are different from and complementary to questions of proximate psychological causation; they are simply at different "levels of analysis" (Tinbergen, 1968; Holekamp & Sherman, 1989).

Evolutionary researchers identify four complementary "levels of analysis" for any behaviour (Tinbergen, 1968; Holekamp & Sherman, 1989). Proximate causes are those that occur within an individual, and include: 1) the specific psychological mechanisms (e.g., emotions) triggered in a situation; and 2) the development of those mechanisms within an individual's lifetime (e.g., learning, gene-environment interactions). Ultimate causes are the reasons that those proximate causes exist at all in any member of the species, and include: 3) the phylogeny or evolutionary history of the psychological mechanisms (e.g., what pre-existing features did the psychological mechanism evolve from); and 4) the evolutionary function (e.g., what selective pressures caused that proximate mechanism to evolve and be maintained despite the costs). These levels are not in competition: to truly understand any phenomenon, we need an answer at all four levels.

[1] Some authors have debated the frequency or effectiveness of punishment (e.g. Guala, 2012), for example by using experiments with unusual methodological constraints that prevent punishers from cooperating (Dreber, Rand, Fudenberg, & Nowak, 2008; Wu et al., 2009). We leave that debate for elsewhere (e.g., see Rankin, dos Santos, & Wedekind, 2009), and instead follow the majority of studies in suggesting that punishment and reward can have some role in supporting cooperation.

These different "levels of analysis" are important to clarify because many of the current debates over cooperation and punishment are due to researchers investigating the same phenomenon at different levels of analysis, and then not realizing that others are simply asking different questions (Barclay, 2012). For example, I may punish Fred because I am angry at his non-cooperation (proximate psychological mechanism), and I may possess this anger because it functions to signal my intolerance of exploitation and thus reduces others' attempts to exploit me (potential ultimate function). These two potential "causes" (anger and signaling) are complementary instead of mutually exclusive, and they could both be right – or both be wrong – because they are answers to questions at two of the four different levels of analysis, namely "what is the specific emotion" and "why does it exist". It is this latter question of ultimate functional cause that we address here. Although we will discuss the costs and benefits of sanctioning, we do not suggest that people are consciously responding to these costs and benefits. Instead, these costs and benefits are what allow the underlying psychological mechanisms (e.g. anger, righteous indignation) to arise. We clarify this now to avoid further confusion over levels of analysis.

What Maintains Peer-to-Peer Sanctioning Behaviour?

Peer-to-peer punishment and rewards can support the provision of public goods, but they themselves are costly to provide. These costs can include time, energy, danger, and especially retaliation (e.g. Janssen & Bushman, 2008; Nikiforakis, 2008). As such, we need to ask why people provide such costly sanctions (Yamagishi, 1986). Sanctions benefit all cooperators in a group, but punishers have a lower payoff than non-punishers in their groups who avoid the cost of punishment. Because of this, non-punishers essentially free-ride on the punishment provided by more punitive group members; this behaviour has been called "second-order free-riding" (where "first-order free-riding" is failing to cooperate). Game theoretical models and evolutionary simulations show that these second-order free-riders will outcompete those who pay to sanction, which then undermines cooperation in those groups because there is no one to enforce it (Boyd & Richerson, 1992; Henrich & Boyd, 2001; Oliver, 1980; Ostrom 1990). A similar argument holds for rewarding: if it is costly to reward cooperators, then rewarders will be outcompeted by non-rewarders, unless some other factor counteracts this disadvantage. Thus, to the extent that punishment and reward maintain cooperation, we need to explain what maintains punitive sentiment and the desire to reward.

Second-Order Sanctions

Second-Order Punishment.

Some theorists have suggested that punitive sentiment is supported by "second-order punishment" or "meta-norms", which means punishment directed preferentially at "second-order

free-riders" (i.e. those who do not punish: Axelrod, 1986; Boyd & Richerson, 1992; Henrich, 2004; Henrich & Boyd, 2001). In these models, "second-order punishment" maintains the "first-order punishment" (i.e. punishment of non-cooperators) which is needed to maintain cooperation. Although this argument appears vulnerable to infinite regress (e.g. "third-order punishment" to maintain second-order punishment, and so on), several models rely on the existence of "second-order" punishment of non-punishers (e.g. Axelrod, 1986; Boyd & Richerson, 1992; Brandt, Hauert, & Sigmund, 2006; Fowler, 2005; Henrich, 2004; Henrich & Boyd, 2001; Sober & Wilson, 1999; de Weerd & Verbrugge, 2011). As such, we need to ask: do people tend to punish non-punishers, as predicted by these models?

The evidence to date suggests that people do *not* perform second-order punishment, i.e. they do not particularly punish non-punishers. In the first analysis we know of, Barclay (2006) examined instances of punishment in a standard public goods experiment and found no unambiguous instances of people punishing a group member for failing to punish in a previous round. Nikiforakis (2008) and Cinyabuguma, Page, & Putterman (2006) [2] both gave experimental participants the opportunity to "counterpunish", i.e. respond to punishment in a second sanctioning opportunity immediately following everyone's first sanctioning decision. They both found that the more a person punishes at the first stage, the more he or she gets punished at the second stage. These two studies suggest that people sanction punishers more than they sanction non-punishers, which is the opposite of what the above models predict. In contrast to those two studies, Denant-Boemont, Masclet, & Noussair (2007) also allowed participants to counterpunish, but they claimed to find evidence for both retaliation and metanorms, even in the very first round of the experiment. In a type of multiple regression, they found that people who punished more than average at the first stage tended to receive more punishment at the second stage (retaliation), and people who punished less than average at the first stage also received more punishment at the second stage (metanorms). Thus, so far one regression analysis claims to find punishment of non-punishers whereas two do not.

In the most direct test to date, Kiyonari and Barclay (2008) had Canadian students play a public goods game, receive an opportunity to sanction cooperators or non-cooperators, and then receive an additional opportunity to sanction partners based on how partners sanctioned. Participants played the game with a free-rider, a cooperator who punished the free-rider, and a cooperator who did not punish. In this direct comparison, non-punishers received no more sanctions than did punishers, and in fact the latter were rated significantly less favourably. This pattern was present

[2] Although Cinyabuguma and colleagues (2006) use the phrase "second-order punishment" in their title, they use it very differently than we use it here. They simply mean a second stage of punishing where people can condition their second-stage punishment on others' first-stage punishment of *anyone*. By contrast, we follow evolutionary researchers in specifically referring to the punishment of those who do not punish free-riders. In their results, Cinyabuguma et al. show that people who punish at the first stage tend to receive more punishment at the second stage, not less, so their results are evidence against – not for – the existence of "second-order punishment" in the sense we use the term here.

in three variations of the experiment, including when the non-punisher was clearly unwilling to use sanctions to support cooperation. Furthermore, these results match similar results in Japan and Belgium (Kiyonari et al., 2004, 2008). These results strongly suggest that people do not preferentially punish non-punishers. As such, these results call into question any model that relies on the existence of second-order punishment to maintain cooperation.

Although there is currently no strong evidence for second-order punishment, at least not in these experimental games, this does not mean it will never be found. There are examples of punishment of non-enforcers in political domains, such as US Senator Joseph McCarthy's accusations of Communism against those who failed to denounce suspected Communists, or the Cuban Liberty and Democratic Solidarity Act of 1996 (a.k.a. the "Helms-Burton Act") which imposes embargoes against companies who fail to embargo Cuba (US Department of State). Both of these examples involve coalitions, which may be relevant, and there are also interesting differences. In the McCarthy example, a failure to sanction was taken as a cue of one's political sympathies; one can similarly imagine "second-order condemnation" of someone who failed to condemn child molestation. Thus, second-order punishment may only arise when these is a clear need to signal one's dislike of another's actions. Alternately, perhaps second-order punishment only arises when an act of non-cooperation is clearly harmful. In the example of the "Helms-Burton Act", one country's failure to sanction undermines the effectiveness of others' sanctions. Future studies can investigate whether second-order punishment will arise when a failure to punish is a potential cue of coalitional membership or future behaviour, and especially when it undermines others' attempts at coercion.

Second-Order Rewards.

Second-order sanctions need not be negative. The rewarding of cooperators could be sustained by "second-order rewards", which means rewards preferentially directed at those who reward cooperators. This would naturally occur if rewarding is seen as a good act within a system of indirect reciprocity or generalized exchange, where people gain a good reputation for helping others and are thus more likely to receive help (see Nowak & Sigmund, 2005 for a review). Panchanathan and Boyd (2004) showed that when the provision of public goods is linked to a system of indirect reciprocity, those who provide public goods become more likely to receive help, and those who refuse to help the public providers will lose their reputation and receive less help themselves. There is no problem of infinite regress because systems of indirect reciprocity can be stable unto themselves (Nowak & Sigmund, 2005). Milinski and colleagues have shown that opportunities for indirect reciprocity can support the provision of public goods (Milinski, Semmann, & Krambeck, 2002), including the fight against climate change (Milinski, Semman, Krambeck, & Marotzke, 2006), as long as people can track reputations about who has contributed and who has not (Semmann, Krambeck, & Milinski, 2004).

Second-order rewards are more likely to support sanctions than is second-order punishment, for several reasons (Kiyonari & Barclay, 2008): a) rewarders will fare better than punishers because the former receive positive reciprocation whereas the latter experience retaliation (e.g. Cinyabuguma et al., 2006); b) punishment requires justification to seem appropriate; c) the same proximate psychological mechanisms are easily co-opted from contributing to rewarding and to second-order rewarding (e.g. liking or empathy towards those who help) whereas this is more difficult with punishment; and d) people will prefer rewarders as partners instead of punishers because the latter can benefit them; and e) rewards and second-order rewards may even constitute a form of competitive altruism (Barclay, 2011a; Barclay & Willer, 2007). In support of this, Kiyonari and Barclay (2008) show that people will readily reward rewarders more than non-rewarders, but do not readily punish non-punishers (or reward punishers).

For rewards to sustain the provision of public goods, provision must be linked to a system of indirect reciprocity (Panchanathan & Boyd, 2004). This means that contributing towards public goods must be seen as good and necessary just as any other act of helping, so people will help public good providers when they are in need. This may require that people recognize that the public good benefits all group members and that failure to contribute harms others (Barclay, 2011b). Once this is common knowledge, good people will be more likely to contribute and no good person will knowingly fail to contribute; this in turn makes reputation systems more effective because rewards and punishment are more likely to go to the appropriate targets. Educating people about such benefits and harms has been successful in the past at promoting reputation systems (Barclay, in press), and experimental evidence shows that such education enhances the effectiveness of indirect reciprocity at promoting public good provision (Milinski et al., 2006).

Unlike punishment, rewards may involve a time lag: it is easy to cheaply inflict immediate high costs on someone for failing to cooperate, but it may be difficult to cheaply reward them with high benefits immediately (b>c) because they may not immediately need help. Humans likely use placeholders such as verbal rewards to acknowledge help received, point out indebtedness, and to indicate that the helper is now more likely to receive tangible help in the future. Verbal rewards can thus substitute as proxy rewards until a cooperator later needs help, and should be effective for those who are able to delay gratification.

Alternative Explanations for Punishment

Punishment as a costly signal.

Punishment may function to signal some characteristic about the punisher (Barclay, 2010). For example, some individuals can punish more cheaply or with lower risk of retaliation, e.g. those with high physical ability or social status can punish more easily (Clutton-Brock & Parker,

1995). Consequently, punishing free-riders could be a socially acceptable way of signaling that one has the power to impose costs on others. Audiences would then defer to a punisher out of self-interest, and this would benefit the punisher. Similarly, punishment could signal one's unwillingness to tolerate exploitation: would-be defectors would benefit from cooperating with anyone known to punish defection. In support of this, game theoretical models show that punishment can evolve when people are less likely to defect on individuals who punish defectors (Brandt, Hauert, & Sigmund, 2003; dos Santos, Rankin, & Wedekind, 2011; Hauert, Haiden, & Sigmund, 2004; Sigmund, Hauert, & Nowak, 2001), and experimental work shows that people steal less money from participants who have punished free-riders (Barclay, submitted). This requires no conscious awareness from either punishers or audiences about punishment's signaling properties: as long as there are reputational benefits, then punitive sentiment will arise and persist.

In addition to being triggered by emotions like anger, punishment of free-riders could also be triggered out of concerns for fairness or righteous indignation (Barclay, 2006). In such cases, anyone who punishes free-riders is demonstrating that he or she possesses a sense of justice and dislikes unfairness. If so, then one could trust a punisher more than someone who does not condemn unfairness. Experiments show that people entrust more money to those who have punished free-riders than to those who have not punished (Barclay, 2006; Nelissen, 2008), though this may require exposure to free-riders so that people understand the justification for punishment. Interestingly, there is currently no evidence that people "like" punishers or will reward them (Barclay, 2006; Horita, 2010; Kiyonari & Barclay, 2008); it seems that people trust punishers (which is in their self-interest) but do not pay to reward them (which is not in their self-interest). This may explain another curious difference in the literature: people pay more to punish others when their decisions are observed than when anonymous (Kurzban, DeScioli, & O'Brien, 2007; Piazza & Bering, 2008), but will hide information about their most severe punishments from an observer who could choose them for future interactions (Rockenbach & Milinski, 2011). Of course, if punishment signals toughness, there is no reason to like punishers or choose them, only to defer to them. More research is needed to determine whether punishment is indeed a signal of some traits, and if so, which ones.

Punishment as a Volunteer's Dilemma.

Researchers often overestimate the costs of punishment because they compare punishers with non-punishers *in the same group*. This overlooks the fact that punishers personally benefit from the cooperation levels they enforce (West, Griffin, Gardner, 2007). As such, punishers can have higher payoffs than non-punishers *in other groups* (i.e. groups without any punishers). Punishment will arise as long as some competition exists against those outside of one's local group, which is to say that competition is not entirely local (see "scale of competition", West et al., 2006).

It is well-established that punishment is a public good (e.g. Yamagishi, 1986). However, not all public goods have the same incentive structure. Most researchers use social dilemmas where non-cooperation is the dominant response (as in a Prisoner's Dilemma), but there are other types of social dilemmas with slightly different incentive structures[3]. Some public goods are a "Volunteer's Dilemma" (Diekmann, 1985; Murnighan, Kim, & Metzger, 1993), also known as a Snowdrift game or Producer-Scrounger game (reviewed by Barclay & Van Vugt, in press). In these situations, each person prefers that someone else provide the public good, but would be willing to provide it if no one else will. For example, the two-person Snowdrift Game models a situation where two people are stuck in cars behind a snowdrift: each person would prefer to stay in his car while the other person shovels the snow, but if the other person refuses to shovel, it is better to shovel alone than to stay stranded behind the snowdrift (Doebeli & Hauert, 2005). This is different from the classic Prisoner's Dilemma because non-cooperation is no longer always the best strategy: it pays to "volunteer" to cooperate if no one else will. Research shows that people are more cooperative in social dilemmas that resemble a Snowdrift Game (2 person) or Volunteer's Dilemma (≥2 person) than in classic Prisoner's Dilemmas (Kümmerli et al., 2007).

Punishment may have an incentive structure that is more like a Volunteer's Dilemma than a Prisoner's Dilemma. Punishment is costly, but all cooperators benefit in the long run if free-riders are punished, and this benefit can outweigh the cost of punishing. As such, each person may prefer that *someone else* be the one to pay to punish a free-rider, but each person might be eventually willing to punish if no one else will (Raihani & Bshary, 2011). Thus, punishment is not always as "altruistic" as is sometimes claimed (e.g. Fehr & Gächter, 2002), but can be self-serving (Bshary & Bshary, 2010; Raihani, Grutter, & Bshary, 2010). Punishment functions as a Volunteer's Dilemma: punishers benefit from their actions, but because they also pay a cost, their net benefit is lower than other group members' (Raihani & Bshary, 2011). Thus, we might expect people to refrain from punishment if it looks like someone else will provide the necessary (and costly) punishment (but see Casari & Luini, 2009).

One way to solve the Volunteer's Dilemma with punishment would be to assign the responsibility for sanctioning to a single individual or specialized role (Baldassari & Grossman, 2011; O'Gorman, Henrich, & Van Vugt, 2009; Ostrom, 1990). By putting the onus on that person(s), it reduces the uncertainty about who would volunteer to sanction, and could even increase the likelihood of sanctions occurring. Alternately, when punishment requires coordination by multiple group members, this can reduce the uncertainty over who will punish and can reduce the costs for everyone (Boyd, Gintis, & Bowles, 2010), as well reduce the frequency with which high contributors get punished (Casari & Luini, 2009).

[3] In fact, an N-Person Prisoner's Dilemma is simply a special case of a more general spectrum of social dilemmas, the rest of which all have a mixed equilibrium that includes cooperators and non-cooperators (Archetti & Scheuring, 2011).

The Carrot or the Stick?

Both punishment and rewards are effective at sustaining cooperation (Fehr & Gächter, 2002; McCusker & Carnevale, 1995; Milinski et al., 2002; Rand et al., 2009; Rockenbach & Milinski, 2006; Sefton et al., 2007; Vyrastekova & van Soest, 2008). When both options are available, people seem to perform both equally at first (Kiyonari & Barclay, 2008), and then use punishment less as cooperation becomes more common and punishment becomes less necessary (Rand et al., 2009; Rockenbach & Milsinki, 2006). Results from social psychology suggest that incentives are more effective when framed as rewards rather than punishments (Komorita, 1987; Komorita & Barth, 1985). Similarly, the psychological literature on learning and operant conditioning suggests that reinforcement-based learning (i.e., reward) is more effective and desirable than punishment-based learning (e.g. Skinner, 1971).

There is debate over whether punishment provides any collective benefit if rewards are also available. Rand and colleagues (2009) gave participants opportunities to pay to reward and/or punish others after each round of a public goods game, and they found that (in comparison with rewards alone) adding punishment had no additional effect on cooperation and in fact reduced collective welfare. However, this lack of difference could be a ceiling effect, as contributions to the public good approached 90% when either rewards or punishment were present. Conversely, Rockenbach and Milinski (2006) paired public goods games with punishment and/or opportunities for indirect reciprocity, and they found that punishment did increase contributions and collective welfare above and beyond the benefits of rewards alone. In fact, they found that participants came to prefer groups with opportunities for costly punishment and indirect reciprocity over groups without punishment (see also Gürerk, Irlenbusch, & Rockenbach, 2006).

It is possible that punishment and reward may be useful for different things. Punishment is cheaper to use when cooperation is common because fewer people need to be sanctioned, whereas rewards are cheaper when cooperation is rare for the same reason (Oliver, 1980). Similarly, punishment can enforce unanimous cooperation by targeting the few rare defectors (Mulder, 2008), whereas reward may be more effective at initiating cooperation from zero by inspiring the few rare cooperators (Forsyth & Hauert, 2011; Hilbe & Sigmund, 2010). If only a few cooperators are required and additional cooperators would be superfluous (e.g. Volunteer's Dilemmas or other situations with a "provision point"), it would be more efficient to use rewards to stimulate a few cooperators rather than punish all defectors into cooperating unnecessarily. Thus, punishment may better when cooperation is common and needs to be unanimous, whereas reward may be better when cooperation is less common or does not need to be. These may all be the reasons why institutions such as governments generally focus on punishment for rare crimes of violence or non-cooperation (e.g. tax evasion, pollution and overharvesting). Similarly,

governments focus on incentives (e.g. tax write-offs) to promote charitable donations and other rarer positive acts rather than criminalize a failure to donate.

Implications and Conclusions

To promote cooperation and the provision of public goods, we need to understand what forces select for and maintain cooperation. It is more effective to alter situations to harness these forces than to focus on directly manipulating proximate psychological mechanisms (Barclay, in press). Because punishment and reward are important in sustaining large-scale cooperation, it is important to understand what forces support the existence of punitive sentiment and a willingness to reward. If our social situations are inconducive to moralistic punishment and rewards via indirect reciprocity, then punishment and reward will likely decrease in frequency and cooperation will subsequently collapse.

Several theoretical models about cooperation rely on non-punishers receiving punishment themselves (e.g. Axelrod, 1986; Boyd & Richerson, 1992; Brandt et al., 2006; Fowler, 2005; Henrich, 2004; Henrich & Boyd, 2001; Sober & Wilson, 1999; de Weerd & Verbrugge, 2011). However, because people do not seem to readily punish non-punishers (Barclay, 2006; Cinyabuguma et al., 2006; Kiyonari & Barclay, 2008; Nikiforakis, 2008; but see Denant-Boemont et al., 2007) we must be hesitant about such models until it is conclusively demonstrated that this "second-order punishment" exists. If it were to be conclusively demonstrated that people do punish those who fail to punish non-cooperators, then we should happily reconsider those models. However, until then, it would be unwise to rely on any model which depends on this highly equivocal phenomenon. Theoreticians and empiricists may wish to investigate other factors that are hypothesized to support the existence of punishment, such as a reputation for punishing (Barclay, 2006; Nelissen, 2008) and especially whether punishment is a Volunteer's Dilemma that exists simply because it pays to punish if no one else is willing to (Raihani & Bshary, 2011).

From a practical perspective, researchers and policy-makers may want to focus on supporting cooperation via means other than just punishment. Rewards and positive reputation have shown to support cooperation in theoretical work (e.g. Barclay, 2011a; Hilbe & Sigmund, 2010; Panchanathan & Boyd, 2004), laboratory work (e.g. Barclay, 2004; Milinski et al., 2002; Rand et al., 2009), and field research (Bateson, Nettle, & Roberts, 2006; Ernest-Jones, Nettle, & Bateson, 2011; Gerber, Green, & Larimer, 2008). Public goods may be provided more effectively if we can design social systems that let people gain positive reputations for doing so.

References

Alexander, R. D. (1987). *The Biology of Moral Systems*. New York: Aldine de Gruyter.

Andreoni, J. (1990). Impure altruism and donations to public goods: a theory of warm-glow giving. *The Economic Journal*, *100*, 464-477.

Archetti, M., & Scheuring, I. (2011). Coexistence of cooperation and defection in public goods games. *Evolution*, *65*, 1140-1148.

Axelrod, R. (1984). *The Evolution of Cooperation*. New York, NY: Basic Books.

Axelrod, R. (1986). An evolutionary approach to norms. *American Political Science Review*, *80*, 1095-1111.

Baldassari, D., & Grossman, G. (2011). Centralized sanctioning and legitimate authority promote cooperation in humans. *Proceedings of the National Academy of Science of the USA*, *108(27)*, 11023-11027.

Barclay, P. (2004). Trustworthiness and Competitive Altruism Can Also Solve the "Tragedy of the Commons". *Evolution & Human Behavior*, *25(4)*, 209-220.

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution & Human Behaviour*, *27*, 344-360.

Barclay, P. (2010). *Reputation and the Evolution of Generous Behavior*. Nova Science Publishers, Hauppauge, NY.

Barclay, P. (2011a). Competitive helping increases with the size of biological markets and invades defection. *Journal of Theoretical Biology*, *281*, 47-55.

Barclay, P. (2011b). The evolution of charitable behaviour and the power of reputation. In C. Roberts (Ed.) *Applied Evolutionary Psychology*, pp. 149-172. Oxford, UK: Oxford University Press.

Barclay, P. (2012). Proximate and ultimate causes of Strong Reciprocity and punishment. *Behavioral and Brain Sciences*, *35(1)*, 16-17.

Barclay, P. (in press). Harnessing the power of reputation: strengths and limits for promoting cooperative behaviours. In press in *Evolutionary Psychology*.

Barclay, P. (submitted). “Don’t mess with the enforcer”: Deterrence as an individual-level benefit for punishing free-riders. Manuscript under review.

Barclay, P., & Van Vugt, M. (in press). The evolutionary psychology of human prosociality: adaptations, mistakes, and byproducts. To appear in D. Schroeder & W. Graziaono (Eds.) *Oxford Handbook of Prosocial Behavior*. Oxford, UK: Oxford University Press.

Barclay, P., & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society of London Series B*, *274*, 749-753.

Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology Letters*, *2*, 412-414.

Batson, C. D., Sager, K., Garst, E., Kang, M., Rubchinsky, K., & Dawson, K. (1997). Is empathy-induced helping due to self-other merging? *Journal of Personality and Social Psychology*, *73*, 495-509.

Boyd, R., Gintis, H., & Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science*, *328*, 617-620.

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

Brandt, H., Hauert, C., & Sigmund, K. (2003). Punishment and reputation in spatial public goods games. *Proceedings of the Royal Society of London Series B, 270*, 1099-1104.

Brandt, H., Hauert, C., & Sigmund, K. (2006). Punishing and abstaining for public goods. *Proceedings of the National Academy of Science of the USA*, *103*, 495-497.

Bshary, A., & Bshary, R. (2010). Self-serving punishment of a common enemy creates a public good in reef fishes. *Current Biology*, *20*, 2032-2035.

Casari, M., & Luini, L. (2009). Cooperation under alternative punishment institutions: an experiment. *Journal of Economic Behavior and Organizations*, *71*, 273-282.

Cialdini, R. B., Brown, S. L., Lewis, B. P., Luce, C., & Neuburg, S. L. (1997). Reinterpreting the empathy-altruism relationship: when one into one equals oneness. *Journal of Personality and Social Psychology*, *73*, 481-494.

Cinyabuguma, M., Page, T. & Putterman, L. (2006). Can second-order punishment deter perverse punishment? *Experimental Economics*, *9*, 265-279.

Clutton-Brock, T. H., & Parker, G. A. (1995). Punishment in animal societies. *Nature*, *373*, 209-216.

Cox, J. C. (2004). How to identify trust and reciprocity. *Games and Economic Behavior*, *46*, 260-281.

Davis, D. D., & Holt, C. A. (1993). *Experimental Economics*. Princeton, NJ: Princeton University Press.

Denant-Boemont, L., Masclet, D., & Noussair, C. N. (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., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science*, *305*, 1254-1258.

Diamond, J. (2005). *Collapse: How Societies Choose to Fail or Succeed*. New York: Viking.

Diekmann, A. (1985). Volunteer's Dilemma. *The Journal of Conflict Resolution, 29*, 605-610.

Doebeli, M., & Hauert, C. (2005). Models of cooperation based on the Prisoner's Dilemma and the Snowdrift game. *Ecology Letters*, *8*, 748-766.

dos Santos, M., Rankin, D. J., & Wedekind, C. (2011). The evolution of punishment through reputation. *Proceedings of the Royal Society of London Series B*, *278*, 371-377.

Dreber, A., Rand, D.G., Fudenberg, D., & Nowak, M. (2008). Winners don't punish. *Nature*, *452*, 348-351.

Ernest-Jones, M., Nettle, D., & Bateson, M. (2011). Effects of eye images on everyday cooperative behavior: a field experiment. *Evolution and Human Behavior*, *32*(3), 172-178.

Fehr, E. & Gächter, S. (2002). Altruistic punishment in humans. *Nature,* 415, 137-140.

Forsyth, P. A. I., & Hauert, C. (2011). Public goods games with reward in finite populations. *Journal of Mathematical Biology*, *63(1)*, 109-123.

Fowler, J. H. (2005). Altruistic punishment and the origin of cooperation. *Proceedings of the National Academy of Science of the USA*, *102*, 7047-7049.

Gächter, S., Renner, E., & Sefton, M. (2008). The long-run benefits of punishment. *Science*, *322*, 1510.

Gerber, A.S., Green, D. P., & Larimer, C. W. (2008). Social pressures and voter turnout: evidence from a large-scale field experiment. *American Political Science Review*, *102(1)*, 33-48.

Guala, F. (2012). Reciprocity: weak or strong? What punishment experiments do (and do not) show. *Behavioral and Brain Sciences*, *35*, 1-59.

Gürerk, O., Irlenbusch, B., & Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science*, *312*, 108-111.

Hardin, G. (1968). The tragedy of the commons. *Science*, 162, 1243-1248.

Hauert, C., Haiden, N., & Sigmund, K. (2004). The dynamics of public goods. *Discrete and Continuous Dynamical Systems B, 4*, 575-587.

Henrich, J. (2004). Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior and Organization, 53*, 3-35.

Henrich, J. & Boyd, R. (2001). Why people punish defectors —Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology, 208*, 79-89.

Hilbe, C., & Sigmund, K. (2010). Incentives and opportunism: from the carrot to the stick. *Proceedings of the Royal Society of London Series B*, *277*, 2427-2433.

Holekamp, K.E., & Sherman, P.W. (1989). Why male ground squirrels disperse: a multilevel analysis explains why only males leave home. *American Scientist*, *77(3)*, 232-239.

Hopfensitz, A., & Reuben, E. (2009). The importance of emotions for the effectiveness of social punishment. *The Economic Journal*, *119*, 1534-1559.

Horita, Y. (2010). Punishers may be chosen as providers but not as recipients. *Letters on Evolutionary Behavioral Science*, *1*, 6-9.

Janssen, M. A., & Bushman, C. (2008). Evolution of cooperation and altruistic punishment when retaliation is possible. *Journal of Theoretical Biology*, *254*, 541-545.

Kiyonari, T., & Barclay, P. (2008). Cooperation in social dilemmas: free-riding may be thwarted by second-order rewards rather than punishment. *Journal of Personality and Social Psychology*, *95(4)*, 826-842.

Kiyonari, T., Declerck, C.H., Boone, C., & Pollet, T. (2008). Does intention matter? A comparison between Public Good and Common Resource Dilemma Games with positive and negative sanctions in one-shot interactions. Paper presentation at the 20th annual meeting of the Human Behavior and Evolution Society, June, 2008, Kyoto University, Kyoto, Japan.

Kiyonari, T., Shimoma, E., & Yamagishi, T. (2004). Second-order punishment in one-shot social dilemma. *International journal of psychology*, 39, 329.

Komorita, S. S. (1987). Cooperative choice in decomposed social dilemmas. *Personality & Social Psychology Bulletin*, *13*, 53-63.

Komorita, S. S., & Barth, J. M. (1985). Components of reward in social dilemmas. *Journal of Personality and Social Psychology*, *48*, 364-373.

Kümmerli, R., Colliard, C., Fietcher, N., Petitpierre, B., Russier, F., & Keller, L. (2007). Human cooperation in social dilemmas: comparing the Snowdrift game with the Prisoner's Dilemma. *Proceedings of the Royal Society of London Series B*, *274*, 2965-2970.

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior*, *28*, 75-84.

Ledyard, J. O. (1995). Public goods: a survey of experimental research. In J. H. Kagel & A. E. Roth (Eds.) *The Handbook of Experimental Economics* (pp. 111-194). Princeton, NJ: Princeton University Press.

Masclet, D., Noussair, C., Tucker, S., & Villeval, M. C. (2003). Monetary and non-monetary punishment in the Voluntary Contributions Mechanism. *The American Economic Review*, *93(1)*, 366-380.

McCusker, C., & Carnevale, P. J. (1995). Framing in resource dilemmas: loss aversion and the moderating effects of sanctions. *Organizational Behavior and Human Decision Processes*, *61*, 190-201.

Milinski, M. Semmann, D., & Krambeck, H. J. (2002). Reputation helps solve the 'tragedy of the commons'. *Nature, 415*, 424-426.

Milinski, M., Semman, D., Krambeck, H. J., & Marotzke, J. (2006). Stabilizing the earth's climate is not a losing game: supporting evidence from public goods experiments. *Proceedings of the National Academy of Science of the USA*, *103*, 3994-3998.

Mulder, L. (2008). The difference between punishment and rewards in fostering moral concerns in social decision making. *Journal of Experimental Social Psychology*, *44*, 1436-1443.

Murnighan, J. K., Kim, J. W., & Metzger, A. R. (1993). The volunteer dilemma. *Administrative Science Quarterly, 38*, 515-538.

Nelissen, R. (2008). The price you pay: cost-dependent reputation effects of altruistic punishment. *Evolution & Human Behavior*, *29(4)*, 242-248.

Nelissen, R., & Zeelenberg, M. (2009). Moral emotions as determinants of third-party punishment: anger, guilt, and the functions of altruistic sanctions. *Judgment and Decision-Making*, *4(7)*, 543-553.

Nikiforakis, N. (2008). Punishment and counter-punishment in public goods games: can we really govern ourselves? *Journal of Public Economics*, *92*, 91-112.

Nikiforakis, N., & Normann, H.T. (2008). A comparative analysis of punishment in public-good experiments. *Experimental Economics*, *11(4)*, 358-369.

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, *437*, 1291-1298.

O'Gorman, R., Henrich, J., & Van Vugt, M. (2009). Constraining free-riding in public goods games: designated solitary punishers can sustain human cooperation. *Proceedings of the Royal Society of London Series B*, *276*, 323-329.

Oliver, P. (1980). Rewards and punishment as selective incentives for collective action: theoretical investigations. *American Journal of Sociology, 85*, 1356-1375.

Ostrom, E. (1990). *Governing the commons.* New York: Cambridge University Press.

Ostrom, E. J., Walker, J. & Gardner, R. (1992). Covenants with and without a sword: self-governance is possible. *American Political Science Review, 86*, 404-417.

Panchanathan, K., & Boyd, R. (2004). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature*, *432*, 499-502.

Piazza, J., & Bering, J. M. (2008). The effects of perceived anonymity on altruistic punishment. *Evolutionary Psychology*, *6(3)*, 487-501.

Raihani, N. J., Grutter, A. S., & Bshary, R. (2010). Punishers benefit from third-party punishment in fish. *Science*, *327*, 171.

Raihani, N. J., & Bshary, R. (2011). The evolution of punishment in n-player public goods games: a volunteer's dilemma. *Evolution*, *65(10)*, 2725-2728.

Rand, D.G., Dreber, A., Ellingsen, T., Fudenberg, D., & Nowak, M.A. (2009). Positive interactions promote cooperation. *Science*, *325*, 1272-1275.

Rankin, D. J., dos Santos, M., & Wedekind, C. (2009). The evolutionary significance of costly punishment is still to be demonstrated. *Proceedings of the National Academy of Science of the USA*, *106(50)*, E135.

Rockenbach, B., & Milinski, M. (2006). The efficient interaction of indirect reciprocity and costly punishment. *Nature*, *444*, 718-723.

Rockenbach, B., & Milinski, M. (2011). To qualify as a social partner, humans hide severe punishment, although their observed cooperativeness is decisive. *Proceedings of the National Academy of Science of the USA*, *108(45)*, 18307-18312.

Sefton, M., Shupp, R., & Walker, J. M. (2007). The effects of rewards and sanctions in provision of public goods. *Economic Inquiry*, *45(4)*, 671-690.

Semmann, D., Krambeck, H.-J., & Milinski, M. (2004). Strategic investment in reputation. *Behavioral Ecology and Sociobiology*, *56*, 248-252.

Sigmund, K., Hauert, C., & Nowak, M. A. (2001). Reward and punishment. *Proceedings of the National Academy of Science of the USA, 98*, 10757-10762.

Skinner, B. F. (1971). *Beyond Freedom and Dignity*. New York, NY: Knopf.

Sober, E. & Wilson, D. S. (1999). *Unto Others: The Evolution and Psychology of Unselfish Behavior* Cambridge, MA: Harvard University Press.

Tinbergen, N. (1968). On war and peace in animals and man. *Science*, *160*, 1411-1418.

Trivers, R.L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, *46*, 35-57.

United States Department of State International Information Programs (n.d.) *Cuban Liberty and Democratic Solidarity (Libertad) Act of 1996.* Retrieved December 13, 2007, from http://usinfo.state.gov/regional/ar/us-cuba/libertad.htm

Vyrastekova, J., & van Soest, D. (2008). On the (in)effectiveness of rewards in sustaining cooperation. *Experimental Economics*, *11*, 53-65.

de Weerd, H., & Verbrugge, R. (2011). Evolution of altruistic punishment in heterogenous populations. *Journal of Theoretical Biology*, *290*, 88-103.

West, S.A., Gardner, A., Shuker, D.M., Reynolds, T., Burton-Chellow, M., Sykes, E.M., Guinnee, M.A., & Griffin, A.S. (2006). Cooperation and the scale of competition in humans. *Current Biology, 16*, 1103-1106.

West, S. A., Griffin, A. S., & Gardner, A. (2007). Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology*, *20*, 415-432.

Wu, J.-J., Zhang, B.-Y., Zhou, Z.X., He, Q.Q., Zheng, X.D., Cressman, R., & Tao, Y. (2009). Costly punishment does not always increase cooperation. *Proceedings of the National Academy of Science of the USA*, *106(41)*, 17448-17551.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology, 51*, 110-116.