J Bioecon (2013) 15:269–279
DOI 10.1007/s10818-013-9153-9

Adaptationist punishment in humans

Robert Kurzban · Peter DeScioli

Published online: 17 February 2013

Abstract Immanuel Kant, Adam Smith, Charles Darwin, George Williams, and Stephen J. Gould, among others, have pointed out that observing that a certain behavior causes a certain effect does not itself license the inference that the effect was the result of intent or design to bring about that effect. Compliance with duty might not reflect the action of conscience, gains in trade might not be due to the benevolence of traders, and fox paws might not be designed to make tracks in snow. Similarly, when person A inflicts costs on person B and, in so doing, generates benefits to C, D, and E (or the group to which A through E belong, in aggregate), the inference that A's imposition of costs on B is by virtue of intent or design to bring about these welfare gains is not logically licensed. In short, labeling punishment "altruistic" because it has the effect of benefitting some individuals is inconsistent with key ideas in philosophy, economics, and biology. Understanding the ultimate cause and proximate design of the mechanisms that cause people to punish is likely to be important for understanding how punishment can help solve collective action problems.

Keywords Punishment · Cooperation · Collective action · emotions · Evolution · Adaptationism

R. Kurzban (✉)
University of Pennsylvania, 3720 Walnut St., Philadelphia, PA 19104, USA
e-mail: kurzban@psych.upenn.edu

R. Kurzban
The University of Alaska Anchorage, Anchorage, AK, USA

P. DeScioli
Stony Brook University, Stony Brook, NY 11794, USA
e-mail: pdescioli@gmail.com

1 Introduction

The relationship between cause and effect is among the most basic issues in science. Some of the most subtle minds in the history of philosophy and science have contributed to this discussion. It's never time ill-spent revisiting past wisdom, and we begin by bringing back to mind well-known quotations that concisely make the point at stake.

Immanuel Kant, in *Groundwork on the Metaphysic of Morals*, wrote:

> In fact there is absolutely no possibility by means of experience to make out with complete certainty a single case in which the maxim of an action that may in other respects conform to duty has rested solely on moral grounds and on the representation of one's duty.

That is, compliance to a moral norm might not be caused by a motive to comply to a moral norm. Fear of punishment or "some secret impulse of self-love," for example, might also account for compliant behavior.

Adam Smith, in *Wealth of Nations*, in arguably the most famous quotation in economics, wrote:

> It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest.

When the baker improves the situation of her customer by providing bread worth more to the buyer than the money given in exchange, beneficence need not be—indeed probably is not—the motive that drives the behavior. The vast majority of welfare improvements in the modern world are side-effects of selfish motives.

Williams (1966) says of a fox having made tracks in the snow,

> On subsequent trips, however, it may follow the same path and have a much easier time of it. . .. This formation of a path through the snow may result in a considerable saving of time and food energy for the fox. . .. Should we therefore regard the paws of a fox as a mechanism for constructing paths through snow? Clearly we should not. (p. 13).

Observing that fox legs have the *effect* of making tracks in the snow does not license the conclusion that they were designed for this purpose. Smith and Kant are interested in proximate motives. Williams is interested in how the details of proximate mechanisms inform our understanding of ultimate function, in this case, what fox paws (mechanisms) were designed to do (functions).

In biology, of course, the means of working through this issue is well understood. By accumulating evidence of design, hypotheses regarding evolutionary causation can be evaluated. Do eyes have design features that make them good for vision? Even closely related functions can often be distinguished. For example, is the Achilles tendon in human legs designed for long-distance running, sprinting, walking, or more than one of these? It turns out that the springy Achilles tendon is one of many aspects of human anatomy showing design for long-distance running (Bramble and Lieberman 2004). To the extent that hypotheses about function predict different design features, it is possible to discriminate among candidate hypotheses (Williams 1966). In psychology,

mechanisms are computational, and hypotheses surrounding function turn on investigation of the information-processing characteristics of these mechanisms (Tooby and Cosmides 1992; Pinker 1997).

Given these ideas, suppose Adam benefits Bob by punishing Charlie, can we infer *either* that Adam was motivated to aid Bob *or* that the mechanism that caused Adam to punish was designed to deliver benefits to third parties? According to the ideas about inferring causes from effects, absent additional evidence, we cannot.

2 Distinguishing altruism from revenge: the evidence

Fehr and Gächter (2002) found that participants in a public goods game in which groups were re-matched each round chose to punish people in their group when given the opportunity to do so, and that people contributed more to the public good when punishment was possible (see Yamagishi 1986; Ostrom et al. 1992 for related findings). The authors concluded that that this "evidence has profound implications for the evolutionary study of human behavior" and that "altruistic punishment is a key force in the establishment of human cooperation."

Recall that when Adam benefits Bob by punishing Charlie, the crux of the claim surrounding altruistic punishment, we cannot infer *either* that Adam was motivated to aid Bob (e.g., Peacock 2007) *or* that the mechanism that caused Adam to punish was designed to deliver benefits to third parties like Charlie. To make this inference, an adaptationist analysis can be brought to bear (Williams 1966). Does the psychology in question have design features that make it organized to execute the putative function, in this case, for delivering benefits to others?

The remainder of this section reviews some evidence relevant to this possibility. As a comparison, we consider the alternative hypothesis that the mechanism causing punishment in these contexts is designed for revenge, imposing costs on those who have imposed costs on (or failed to confer benefits on) oneself (Lieberman and Linke 2007; McCullough 2008; McCullough et al. 2013b). Our claim is not necessarily that these systems are, in fact, designed for revenge, but rather that the adaptationist approach is useful to differentiate among candidate hypotheses given the available data, and it is possible to compare the known design features against the predictions of an altruism model and a revenge model.

The strength of this approach is that adaptationist analysis avoids the error of inferring design or intent from effects because of the clear evidentiary criteria established by Williams (1966) and the "onerous" burden assumed by such an analysis. Hypothesized biological functions make commitments with respect to the properties of the mechanism—physiological or psychological (i.e., computational)—under study. That is, to propose a biological function is to commit to the view that the mechanism in question has design features that make it improbably well organized to execute the putative function. Measurements and experiments can then be used to support or undermine functional claims. If empirical patterns do not provide evidence of the predicted design features—as in the springiness that is predicted by a long-distance-running function for legs in the example above—then confidence in the proposed function should be diminished.

In short, to draw conclusions about design, looking at the *effect* of Adam's (punishment) behavior is not sufficient and additional evidence is required beyond the effect of punishment, per the guidance of our philosophers. A first step of biological analysis is articulating the design features each model predicts and evaluating the evidence and data against these predictions.

Roughly, revenge functions to deter others from subsequent imposition of costs, or withholding of benefits, from ego or ego's allies (McCullough et al. 2013a).[1] This deterrence function is implemented by imposing costs on those who have intentionally imposed costs on ego, deterring subsequent imposition of costs by the same individual or others. Hence, revenge can be useful in deterring other people from behaving similarly in future interactions. So, if the revenge system is designed for a world in which interactions are generally repeated and interactions are observed, revenge should be expected even when additional interactions with the person who harmed the actor are excluded, such as the last round of experimental games (Burnham and Johnson 2005; Hagen and Hammerstein 2006). That is, the input activating the system is taken to be past harm, and the output is the motive to impose costs on the agent in question.

The relevant design features implied by a putative revenge system lead us to ask whether the mechanism causes individuals to (1) punish if and only if ego (or ego's allies) were harmed, (2) punish as a function of the deviation from the relevant baseline of benefits to which one is entitled, (3) punish even in the absence of subsequent interactions, and (4) experience the emotion of anger, the affective system typically associated with motivating revenge. Humans take revenge under many circumstances, and some have argued it plays an important role in economic games: Falk et al. (2005), for instance, concluded that in some contexts, "retaliation, i.e., the desire to harm those who committed unfair acts, seems to be the most important motive…" (p. 2017).

In contrast, a hypothesized altruism system (that punishes to benefit someone else) implies four straightforward and parallel properties: the system should cause individuals to (1) punish even if ego (or ego's allies) were not harmed, (2) punish as a function of the absolute level of harm, (3) punish if and only if the punishment will alter the behavior of the individual being punished in subsequent interactions,[2] and (4) experience the emotion of empathy, the affective system strongly associated with motivating altruism (Batson et al. 2002). In addition, if punishment is designed for altruism, then one might expect that punishment generally leads to improved aggregate outcomes in groups in which punishment is possible relative to those in which it is not.

Before turning to the evidence, we pause for important terminological points. Fehr and Gächter (2002) define "altruistic punishment" behaviorally, taking it to mean that "individuals punish, although the punishment is costly for them and yields no material gain" (p. 137). According to the above ideas, an individual's punishment behavior might meet the criteria of this behavioral definition—being costly and yielding

[1] For brevity, only costs, rather than the withholding of benefits, is used hereafter though the logic is the same because preventing a cost and causing a benefit are equivalent for this analysis.

[2] There seems to be disagreement on this point. The logic of punishing to induce subsequent altruism seems to imply the need for subsequent opportunities for altruism. Hence, this idea seems to predict no end-game punishment.

no gain—while simultaneously failing to be the result of mechanisms designed for altruism. This highlights the problem with behavioral definitions raised above. Indeed, what researchers call "anti-social punishment" (Hermann et al. 2008) also meets the above definition of "altruistic punishment"—costly and without benefit to the punisher—illustrating that the proposed definition is problematic.

So, we wish to be clear that our claim is not that punishment in public goods games is not "altruistic punishment" *as Fehr and Gächter (2002) defined that term,* but rather that behavior that is so classified under such a definition might have been caused by mechanisms other than benefit-delivery systems. In particular, we are proposing that "altruistic punishment" (again, as defined by Fehr and Gächter 2002) might be caused by systems designed for deterrence (revenge) rather than altruism.

We also note that this punishment might fit a biological behavioral definition of altruism—raising another organism's lifetime reproductive success while reducing the lifetime reproductive success of the actor (Hamilton 1964)—while, again, arising from mechanisms designed for a function other than benefit-delivery. Similar to the present argument, the limitations of definitions only in terms of behavior have been recognized by West et al. (2011), who argue for including not just the effect of behavior but also its function when they define cooperation as "a behaviour which provides a benefit to another individual (recipient) *and which is selected for because of its beneficial effect on the recipient*" (p. 235, emphasis added). To illustrate with an example, they write that "when an elephant produces dung, this is beneficial to the elephant (emptying waste) and also beneficial to a dung beetle that comes along and uses that dung, but it is not useful to call this cooperation." (p. 235). The logic we apply to punishment is directly parallel, asking if there is evidence that the mechanisms causing punishment might be selected for one function, but producing benefits to others as a side-effect, in the same way that elephant dung aids dung beetles as a side-effect. (See also West et al. 2007.)

We now turn to evidence relevant to arbitrating among these different proposals.

Line 1: Who is Punished?

The most obvious line of evidence relevant to distinguishing revenge from altruistic punishment is the issue of harm to the potential punisher. This is a crucial feature of revenge, but irrelevant to a mechanism designed to benefit others.

Arguably the cleanest investigation of this is the one-shot games Carpenter and Matthews (2012) conducted which varied who subjects could punish: members of one's own group or members of another group. In the key treatment, in which members of one group could punish individuals in another, but not vice-versa (which opens up the possibility of between-group reciprocal punishment), they find that ten percent of participants punished individuals in the other group, averaging about $0.10—which might still include some punishment due to confusion (Andreoni 1995; Kurzban and Houser 2001) and punishment due to spite. This implies that punishment in public goods games *absent revenge* could be indistinguishable from zero.

These data strongly undermine the view that punishment is driven by the motive for altruism, and supports the view that punishment in public goods games is driven by the motive for revenge.

Line 2: What is Punished?

A mechanism designed to elicit altruism should function to straightforwardly incentivize the delivery of benefits to others. Punishment, or the threat of it, makes altruism a less costly choice than selfishness. In the linear public goods environment, every incremental increase in contribution leads to additional benefits to everyone in the group. So, from the point of view of helping others, *there is nothing special, at all, about a potential punisher's own contribution or the average contribution of the group*. The goal of an altruistic punisher is to punish sufficiently to induce maximum contributions. If a player contributes the maximum minus X, an altruistic punisher should reduce that player's payoff a little more than the marginal value of keeping X. If the player takes this into account in the next round, they will contribute the maximum. Altruistic punishers should, then, punish as a linear function of deviation from the maximum possible contribution.

Revenge, in contrast, is not designed around maximizing benefits to others. If ego contributes little to a public good, it does not make sense to talk about ego "taking revenge" on someone who did not contribute the maximum. To the extent that public goods contributions are viewed as reciprocal obligations (e.g., Kurzban et al. 2001), revenge should be taken with respect to deviation from fulfillment of these perceived obligations rather than deviation from the maximum.

So, the two models make different predictions about the relationship between punishment and the cooperativeness of the players in public goods games. Altruistic punishment systems should punish as a function of deviation from the maximum, and there should be nothing special about the point of average contribution to the group as long as this average is below maximum possible contributions. In contrast, to the extent contributions in public goods games are construed by players in the context of reciprocal obligations, revenge should increase as a function of deviation from the relevant baseline, the average cooperativeness of group members.

The data from Fehr and Gächter (2002) show the functional relationship predicted by the revenge model, with an abrupt change at the mean contribution. Though there is variability, the bulk of punishment cross-culturally is similarly directed at those who contribute less than the punisher (Hermann et al. 2008). Subsequent work has shown that the difference between a subject's contribution and that of the individual they are punishing strongly predicts punishment decisions (O'Gorman et al. 2008; Cubitt et al. 2011). These patterns of data undermine the view that punishment is driven by the motive for altruism.

Line 3: Are Subjects Angry or Empathic?

The signature emotion associate with altruism is empathy (Batson 1991), and a vast corpus of research spanning many decades, many cultures, and an array of research methods converges on this conclusion (e.g., Batson et al. 2002). Revenge, on the other hand, is associated with anger, the emotion that seems to be the motivational system that drives people to impose costs on those who have imposed costs (or withheld benefits) from oneself (McCullough 2008).

Fehr and Gächter (2002) show that subjects experience anger when they think about those who contribute little to a public good relative to others. (See also Cubitt et al. 2011.) This is easily consistent with a revenge interpretation. Peacock (2007) takes a

strong view, saying that it "is not immediately plausible that punishment be altruistic if, for instance, the punisher's motive is to harm or vent his anger on another person" (p. 11). If the design of the mechanism in question is to benefit others, then the primary emotion should be expected to be empathy for those adversely affected, with anger perhaps as a secondary emotional experience. The current evidence shows the reverse and undermines the altruistic punishment hypothesis.

Line 4: Is there a Future?

Fehr and Gächter (2002) say that "punishment may well benefit the future group members of a punished subject, if that subject responds to the punishment by raising investments in the following periods" and that it is "*in this sense*" (our italics) that "punishment is altruistic" (p. 137).[3] This seems to commit them to the view that the punishment they are interested in, to be altruistic, must benefit future group members of the subject who is punished. Punishment in the final round of a multi-round game, or in one-shot games, clearly cannot be altruistic *in that sense*. As discussed above, revenge interpretations are not so committed.

Empirically, punishment does not decline in the last round of repeated public goods games (Fehr and Gächter 2002; Anderson and Putterman 2006; Gächter et al. 2008) and is observed in one-shot games (Halloran and Walker 2004), strongly undermining the view that punishment is driven by the motive for altruism given the above conceptual commitment. These findings make sense, however, in the context of a proximate mechanism of revenge, designed for a world with many repeat interactions (McCullough et al. 2013a).

Line 5: Who, if anyone, Benefits?

Finally, suppose we ignore Adam Smith's logic and judge acts altruistic just in case they make people better off. Does punishment improve welfare in PG games? Dreber et al. (2008) reported: "punishment does not increase the average payoff. In some experiments, punishment reduces the average payoff, whereas in others it does not lead to a significant change. Only once has punishment been found to increase the average payoff" (p. 350).[4] (See also Gächter et al. 2008, for an additional exception and Gächter and Herrmann 2009, for a discussion.) The size of the negative effect of punishment on earnings is occasionally quite striking in size (O'Gorman et al. 2008).

That is, *even if* one makes the mistake Adam Smith and Kant warned about, and we refer to behavior as altruistic if it has the *effect* of making people better off, behavior in Fehr and Gächter's work and most (but not all) subsequent public goods games would *still* not be altruistic since the relevant agents are on average worse off.

[3] Use of the term is inconsistent, and other users (e.g., Boyd et al., 2003) are not so committed. Oddly, in one article investigating "The Neural Basis of Altruistic Punishment," (de Quervain et al. 2004), the phenomenon as defined is *ruled out*, since there is no play following punishment.

[4] The exception here is to (Nikiforakis and Normann 2008). Ostrom et al. (1992) found that, with communication, the possibility of punishment improved outcomes a great deal in a social dilemma, though this was not the case without communication.

2.1 Summary

Five lines of evidence reviewed briefly here converge on the view that the data, as they currently stand, weigh heavily against the view that punishment in public goods games is driven by a psychological mechanism designed for altruism.[5]

3 Evidence from other games

A thorough treatment of research outside of public goods games that potentially speak to the issue of the extent to which people punish third parties for behavior in which they themselves are not involved is beyond the scope of the present work. A few stylized findings are briefly reviewed here.

First, Fehr and Fischbacher (2004) found that third parties spend a small amount (on average US$1, or 3.35 out of 20 points possible for this purpose) to punish those who defect against a cooperative partner in a PD game. However, they find that "sanctions by second parties directly harmed were much stronger than third-party sanctions" (p. 85). Marlowe et al. (2008) showed in a cross-cultural sample (Henrich et al. 2006) that subjects punish those who give relatively little in a Dictator Game. Taken together, these two results illustrate that third party punishment occurs both when there are aggregate welfare losses and when there are not, raising the possibility that third party punishment might have little to do with inducing others to deliver benefits.

Second, third party punishment (unlike revenge; see Bolton and Zwick 1995) diminishes when such acts are carefully kept anonymous (Kurzban et al. 2007; Piazza and Bering 2008), in which case the modal behavior is no punishment at all. In addition, punishment is occasionally "perverse" (Cinyabuguma et al. 2006), with those who are cooperative, fair, or generous suffering sanctions. These findings highlight that measuring punishment requires controls in which people can punish those who cooperate.

Finally, the revenge interpretation predicts that one's allies and friends will be avenged. Recently, Bernhard et al. (2006) found that "punishers protect ingroup victims…much more than they do outgroup victims" (p. 912), providing evidence in favor of a revenge interpretation.

4 Concluding remarks

Is there "altruistic punishment in humans?" In one sense, yes. Fehr and Gächter's (2002) definition—"individuals punish, although the punishment is costly for them and yields no material gain"—includes an enormous number of behaviors, including cases of spite and revenge. (See also Peacock 2007, on this point.) Defining X as altruistic if X yields no material gain (holding aside the fact that the definition does not require that anyone benefits) gives rise to the problem Adam Smith warned against.

[5] This of course does not decide the issue. Price et al. (2002) would likely argue the data are consistent with their "fitness-leveling" account.

Paying to see a movie is "altruistic" in this sense because the viewer incurs a cost but reaps no material gain.[6]

In some sense, these definitional issues are irrelevant, since Fehr and Gächter (2002), and others, are probably not interested in "altruistic punishment" as they defined it. They are interested in punishment as a possible explanation for cooperation among non-kin human groups. That is, they are interested in the generation of aggregate benefits, not the possibility that people impose costs on others at a cost to themselves *per se*.

The disconnect between the definition and the real issues at stake again derives from the behavioral definition. These definitions run into difficulties because they do not respect the issue raised by Kant, Smith, and Williams. Observing that some individuals benefit from some act does not, by itself, license the inference that these benefits were the proximate goal of the act or that the phenotype that produced the benefits was designed to do so.

In sum, there are (at least) two possibilities. One is that humans have cognitive systems that motivate them to punish others, even if they themselves were not previously harmed, in order to bring about benefits to third parties. In certain respects, this possibility leads to a great deal of optimism for institutional design, since institutions could be structured to let people carry out their wishes and punish potential non-cooperators. The second possibility, not mutually exclusive with the first, is that humans have cognitive systems that motivate them to punish those who previously harmed them (relative to baseline expectations). If this is the case, then in certain circumstances, punishment so motivated can, as a side-effect, happen to lead to aggregate benefits when there is a tight relationship between the punisher's payoffs and others' payoffs. This possibility, if true, points to very different institutional solutions.

Distinguishing between these two possibilities has already come a substantial way, and the evidence weighs against the view that punishment is motivated by a system designed to deliver benefits to others. Revenge remains a viable alternative. It might, of course, be that humans bring to bear multiple motives in these contexts, and the cross-cultural covariance of altruism and punishment in experimental games is noteworthy (Henrich et al. 2006). Additional work will be needed to explore the relative magnitudes of revenge and other motives in various contexts.

On a broader level, the philosophical commitments of adaptationism are useful in clarifying the issues at stake and the associated basis for evaluating evidence in biology and its sub-discipline, psychology. Continued resistance to using adaptationism in theorizing and defining terms of art is puzzling given the clarity its use affords and the rigor it imposes in drawing inferences from data. Indeed, adaptationist reasoning and adaptationist definitions are absolutely essential for continued progress at the intersection of biology and social science.

[6] This example is not intended to be flip, but to point out the logical entailments of the definition. It is worth noting that the psychological gain cannot be considered to offset the cost of the movie, making the transaction a net positive for the movie-goer, since these considerations would also then need to be applied to the pleasure of punishing. This consideration, presumably, underlies the choice of the word "material" in the definition.

References

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

Andreoni, J. (1995). Cooperation in public goods experiments: Kindness or confusion? *American Economic Review*, *85*, 891–904.

Batson, C. D. (1991). *The altruism question: Toward a social-psychological answer*. Hillsdale, NJ: Erlbaum.

Batson, C. D., Ahmad, N., Lishner, D. A., & Tsang, J. (2002). Empathy and altruism. In C. R. Snyder & S. L. Lopez (Eds.), *Handbook of positive psychology* (pp. 485–498). New York: Oxford University Press.

Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature*, *442*(24), 912–915.

Bolton, G. E., & Zwick, R. (1995). Anonymity versus punishment in ultimatum bargaining. *Games and Economic Behavior*, *10*, 95–121.

Bramble, D. M., & Lieberman, D. E. (2004). Endurance running and the evolution of Homo. *Nature*, *432*, 345–352.

Burnham, T., & Johnson, D. D. P. (2005). The evolutionary and biological logic of human cooperation. *Analyse & Kritik* (Special issue on Ernst Fehr), *27*(1), 113–135.

Carpenter, J., & Matthews, P. (2012). Norm enforcement: Anger, indignation, or reciprocity. *Journal of the European Economic Association*, *10*(3), 555–572.

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

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

de Quervain, D., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., et al. (2004). The neural basis of altruistic punishment. *Science*, *305*, 1254–8.

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

Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving forces behind informal sanctions. *Econometrica*, *7*, 2017–2030.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*, 63–87.

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

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

Gächter, S., & Herrmann, B. (2009). Reciprocity, culture and human cooperation: Previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1518), 791–806.

Gächter, S., Renner, E., & Sefton, M. (2008). The long-run benefits of punishment. *Science*, *322*(5907), 1510–1510.

Hagen, E. H., & Hammerstein, P. (2006). Game theory and human evolution: A critique of some recent interpretations of experimental games. *Theoretical Population Biology*, *69*, 339–348.

Halloran, M. A., & Walker, J. M. (2004). Rewards and sanctions and the provision of public goods in one-shot settings. *Experimental Economics*, *7*, 235–247.

Hamilton, W. D. (1964). The genetical evolution of social behaviour, I & II. *Journal of Theoretical Biology*, *7*, 1–52.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanataz, A., et al. (2006). Costly punishment across human societies. *Science*, *312*, 1767–1770.

Hermann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, *319*, 1362–1367.

Kurzban, R., & Houser, D. (2001). Individual differences and cooperation in a circular public goods game. *European Journal of Personality*, *15*, S37–S52.

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

Kurzban, R., McCabe, K., Smith, V. L., & Wilson, B. J. (2001). Incremental commitment and reciprocity in a real time public goods game. *Personality and Social Psychology Bulletin*, *27*, 1662–1673.

Lieberman, D., & Linke, L. (2007). The effect of social category on third party punishment. *Evolutionary Psychology*, *5*, 289–305.

Marlowe, F. W., Berbesque, J. C., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Ensminger, J., Gurven, M., Gwako, E., Henrich, J., Henrich, N., Lesorogol, C., McElreath, R., & Tracer, D. (2008). More 'altruistic' punishment in larger societies. *Proceedings of the Royal Society Biology*, *275*, 587–590.

McCullough, M. (2008). *Beyond revenge: The evolution of the forgiveness instinct*. San Francisco: Jossey-Bass.

McCullough, M. E., Kurzban, R., & Tabak, B. A. (2013a). Cognitive systems for revenge and forgiveness. *Behavioral & Brain Sciences, 36*, 1–15.

McCullough, M. E., Kurzban, R., & Tabak, B. A. (2013b). Revenge, forgiveness, and evolution. In M. Mikulincer & P. R. Shaver (Eds.), *Understanding and reducing aggression, violence, and their consequences*. Washington, DC: American Psychological Association.

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

O'Gorman, R., Henrich, J., & Van Vugt, M. (2008). Constraining free-riding in public goods games: Designated solitary punishers can sustain human cooperation. *Proceedings of the Royal Society B*, *276*, 323–329.

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

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

Peacock, M. S. (2007). The conceptual construction of altruism: Ernst Fehr's experimental approach to human conduct. *Philosophy of the Social Sciences*, *37*, 3–23.

Pinker, S. (1997). *How the mind works*. New York, NY: W. W. Norton & Company.

Price, M. E., Cosmides, L., & Tooby, J. (2002). Punitive sentiment as an anti-free rider psychological device. *Evolution and Human Behavior*, *23*, 203–231.

Smith, A. (1776). *Wealth of Nations*. London: W. Strahan and T. Cadel.

Tooby, J., & Cosmides, L. (1992). Psychological foundations of culture. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 19–136). New York: Oxford University Press.

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.

West, S. A., Mouden, C. E., & Gardner, A. (2011). Sixteen common misconceptions about the evolution of cooperation in humans. *Evolution and Human Behavior*, *32*, 231–262.

Williams, G. C. (1966). *Adaptation and natural selection: A critique of some current evolutionary thought*. Princeton: Princeton Press.

Toshio, Y. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, *51*, 110–116.