On Hotheads and Dirty Harries

The Primacy of Anger in Altruistic Punishment

Elise C. Seip,[a] Wilco W. van Dijk,[b] and Mark Rotteveel[a]

[a]*University of Amsterdam, Amsterdam, the Netherlands*

[b]*VU University Amsterdam, Amsterdam, the Netherlands*

Recent research has shown that individuals are prepared to incur costs to punish non-cooperators, even in one-shot interactions. However, why would people punish non-cooperators with no apparent benefits for the punishers themselves? This behavior is also known as altruistic punishment. When defection is discovered, an individual evaluates this act as unfair, which could result in anger. We argue that although unfairness and anger are often intertwined, it is primarily the experience of anger and not the perception of unfairness that produces altruistic punishment. We briefly present recent data in line with the hypothesis that identifies anger as the underlying mechanism of altruistic punishment. Furthermore, additional influences regarding the occurrence of altruistic punishment, e.g., intentionality of the interaction partner, the role of satisfaction, and individual differences, are discussed.

Key words: **altruistic punishment; anger; fairness; emotion**

Altruistic Punishment

Scholars from both biological and social sciences face the long-standing problem of understanding the conditions required for the emergence and maintenance of human cooperation. Unlike other organisms, people frequently cooperate with genetically unrelated strangers, with people they never meet again, and when reputation gains are small or even absent. This behavior is puzzling because such cooperation can incur individual costs to confer benefits on unrelated others. It cannot be explained by mechanisms commonly used to explain cooperative behavior in general, e.g., nepotistic motives associated with the theory of kin selection and inclusive fitness,[1] selfish motives associated with theories of direct reciprocity,[2,3] indirect reciprocity based on reputation,[4,5] and costly signaling.[6] Recently it has been suggested that altruistic punishment, that is, people's propensity to incur costs in order to punish non-cooperators, might provide one solution to this intriguing puzzle of human cooperation.[7–10] Altruistic punishment[a] means that individuals punish non-cooperators (i.e., *free riders or defectors*) even if this punishment is costly (e.g., in terms of money, time, or effort) and yields no apparent benefits for the punishers themselves.[10] However, punishment may well benefit future interaction partners of the punishee if the punishee responds to the punishment by increasing cooperation in future interactions. Indeed, it has been shown that people do engage in altruistic punishment

Address for correspondence: Elise C. Seip, MSc., University of Amsterdam, Psychology Department, Room 906, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands. Voice: +31 20 5256292; fax: 020 639 1896. E.C.Seip@uva.nl

[a]Some scholars use the term *costly punishment* instead of altruistic punishment. Although the latter term is most often used in the literature, the former is in our view more precise. One might argue whether altruistic punishment, as described in the literature, is always truly altruistic. For instance, punishment can be regarded as self-interested, despite its private costs, if the punisher benefits from increased public good provision over the long term[37] or if, as we argue, the punisher derives satisfaction from imposing punishment.

Values, Empathy, and Fairness across Social Barriers: Ann. N.Y. Acad. Sci. 1167: 190–196 (2009).
doi: 10.1111/j.1749-6632.2009.04503.x

even in one-shot interactions and that cooperation flourishes if altruistic punishment is possible and breaks down if it is ruled out.[9,10] Moreover, anthropological research has yielded results from 15 populations on five continents showing that all studied populations engaged in altruistic punishment and that its magnitude covaries positively with altruistic behavior across these populations.[11] However, care should be taken in generalizing these findings because recent research showed important cross-societal differences in punishment and cooperation;[12] in some societies cooperation did not increase upon presence of punishment. Punishment seems only to enhance cooperation in the presence of strong social norms.

Although altruistic punishment may represent an attractive and plausible solution for the puzzle of human cooperation (at least in strong social norm societies), it also creates a new puzzle that calls for further theoretical and empirical investigation. Why would people incur costs to punish free riders and in this way provide benefits to (unrelated) others in the first place? Taking an inconsiderate clod to task for butting into line in front of you makes perfect sense, but how does one explain the person who bawls out a stranger for butting into line *behind* him? Because punishment is costly for the individual but beneficial for the group as a whole, it creates a second-order social dilemma.[13] Everyone in a population will be better off if noncooperation (and norm violation) is deterred, but nobody seems to have an individual incentive to bear the costs of punishing defectors. In the present article we argue that to solve this dilemma, at least partly, a closer look at the emotional processes underlying altruistic punishment is needed.

Emotions and Altruistic Punishment

Emotions have evolved to prepare us to address important events in our lives adaptively. The experience of an emotion can be considered as a felt action tendency[14,15] but also guides specific action and information processing tendencies.[16] For example, fear can be characterized by tendencies to avoid and prepare for vigorous action and to carefully scrutinize the surroundings for signs of actual threat. Positive affect, such as in happiness or joy, on the other hand broadens our attention,[17] evokes playful behavior, and facilitates social interaction.[18] In the case of anger, an approach action tendency is initiated to remove the obstacle (the situation is worth fighting for) accompanied by high alertness.[14,19] In other words, emotions exist for the regulation of oneself and one's behavior in relation to another person, an event, or even an object.[14,15,20] We argue that emotions may contribute to an optimal solution for the individual as well as for the group by evoking altruistic punishment and subsequent cooperation. The experience of anger and subsequent punishment might be a way to express anger, and at the same time punishment might be a way to communicate a social norm, promoting cooperation in the end.

In the present article we argue that specific emotions constitute the proximate mechanism underlying altruistic punishment. Emotions can direct punishment behavior in spite of demands of effort or other costs. More specifically, we argue that the experience of anger can provide "extra fuel" necessary for people to punish free riders even if it is costly and does not yield any material benefits. Theorists from different disciplines have argued that anger can be triggered by violations of reciprocity. If defection is discovered, an individual appraises this act as unfair and blameworthy and, consequently, experiences anger toward the defector.[21,22] Former research has already shown that altruistic punishment in response to unfair behavior is related to feelings of anger toward the defector.[10] In contrast, fair offers are related to feelings of happiness and activation in the neural reward circuitry.[23]

If anger constitutes a proximate cause of altruistic punishment, one would expect that

anger will be more intense as the contributions of free riders deviate further from average investments. This line of reasoning is supported indirectly by research showing that unfair (low) offers in economic games are accompanied by an increase in arousal (a proximate for emotion) that itself is correlated positively with the subsequent rejection of these unfair offers.[24] In this study by van 't Wout and colleagues, no additional measures of anger were assessed, unfortunately, therefore we consider this as indirect evidence. More indirect evidence for the hypothesis that anger causes altruistic punishment comes from recent (neuro) economic studies. Sanfey and colleagues[25] showed, for example, that unfair offers in the ultimatum game elicited activity in the anterior insula, a brain area that is also associated with anger. Moreover, they showed that this brain activity was correlated with the subsequent decision to reject the unfair offer (a proxy for altruistic punishment). No evidence was obtained, however, regarding subjective feelings of anger that could have led to rejections of unfair offers. Also, areas associated with reward processing, such as the orbitofrontal cortex and nucleus accumbens, were activated (in men but not in women) upon a cue indicating high shock to unfair players. This activity even correlated with an expressed desire for revenge and was accompanied by a decrease in empathic neural response.[26] Here it is not clear, however, whether the men were prepared to incur costs to punish the unfair player; there was also no direct measurement of the perception of unfairness and anger. Therefore these findings provide only indirect evidence for the role of anger in altruistic punishment.

In summary, these studies suggest the involvement of emotion and, more specifically, anger in altruistic punishment. How exactly the unfairness of the situation is related to anger and subsequently to the use of altruistic punishment remains, however, open to question. In the following paragraphs we provide additional evidence and give a suggestion for the interplay between anger, unfairness, and altruistic punishment.

Perception of Unfairness and the Experience of Anger

One important and necessary component of the experience of anger could be the perception of unfairness.[27] We argue that although unfairness and anger are often intertwined, it is the experience of anger and not the perceived unfairness that produces altruistic punishment. In dealing with an unfair situation, the dynamical interaction of cognitive-emotional systems might result in two types of emotional control processes (direct as well as indirect), as proposed by Ochsner and colleagues,[28] disentangling anger on the one hand and perception of unfairness on the other. The first process is related to onset and experience of anger where the context-appropriate emotional value of a stimulus is evaluated, like the unfairness of an offer. Consequently, actions based upon these evaluations are selected, like the decision to punish. This control process contains direct reciprocal connections between ventral prefrontal cortex and orbitofrontal systems. The second type might be related only to the perception of unfairness. Here the association between stimulus and emotional responses can explicitly be described and reasoned about. This could lead to an alternative evaluation of the event and might modulate the direct emotional reaction. The evaluation of unfairness might influence the occurrence of anger and punishment. This second process involves dorsal prefrontal systems that have few, if any, connections with emotional appraisal systems.

Evidence that the perception of unfairness is the result of explicit reasoning comes from a study by Knoch and colleagues.[29] Disruption of the right, but not left, dorsolateral prefrontal cortex (DLPFC) by transcranial magnetic stimulation (TMS) substantially reduced subjects' willingness to reject their partners' intentionally unfair offer. This suggests that subjects were less able to resist the economic temptation to accept these offers. In addition, upon applying TMS, subjects accepted unfair offers almost as quickly as fair offers, suggesting that, with

disruption of DLPFC, self-interest impulses have a stronger impact on behavior.[29] Rejection of an unfair offer might, therefore, involve a more controlled higher order process. These findings suggest that punishment behavior is related to perception of unfairness. We agree with the importance of unfairness perception in punishment behavior but we also argue that it is not the perception that drives punishment behavior but the experience of anger that triggers altruistic punishment. Below we will describe some of our recent work that is in line with our reasoning and that provides direct empirical support for the primary role of anger in altruistic punishment.

Anger as a Proximate Mechanism for Altruistic Punishment

Pillutla and colleagues[27] were the first to study the relationship between the role of perceived fairness and emotions in the decision to either reject or accept unfair (low) offers. They showed that anger in response to an unfair offer was a better predictor of rejections than the perception of unfairness per se. Results were only correlational, therefore no statements could be made regarding the causal relationship between anger and altruistic punishment. In addition, Pillutla and colleagues[27] used the ultimatum game in which participants only had the choice to accept or reject passively an offer. We used a paradigm in which participants could act directly toward the defector with altruistic punishment. Since anger is reflected more by action, such as the tendency to approach the defector,[19] whereas there is not at all an action tendency in the decision to accept or reject an offer, we think that altruistic punishment can be better studied within a trust paradigm than an ultimatum game.

In a series of three studies, we examined the impact of both unfairness and anger on altruistic punishment. In our first study we measured participants' reactions to a non-cooperator in a sequential trust game. Results showed that the more non-cooperation was perceived as unfair, the more anger it evoked and the more punishment was given to the non-cooperator. Importantly, follow-up analyses showed that the impact of unfairness on punishment was fully mediated by experienced anger. That is, perceived unfairness elicits anger and, subsequently, this anger triggers punishment. These results indicate that perceived unfairness per se does not predict punishment but that punishment is more reliably predicted by feelings of anger.

In a second study we investigated the role of anger in altruistic punishment. In this study participants played a one-shot, three-person, public good game 10 times. Results showed that the lower the contribution of others, the more angry participants felt and the harsher they punished others. In line with our reasoning, follow-up analyses showed that the effect of the contribution of others on imposed punishment was fully mediated by experienced anger. These results suggest again that punishment is not triggered by the others' contributions to the public good but by people's anger. Thus, both studies indicate that anger and not the perception of unfairness is a strong and reliable predictor of altruistic punishment.

To further test whether it is primarily anger that underlies altruistic punishment, we primed participants with feelings of anger in our third study. Half of the participants were asked to recall and describe an autobiographical episode in which they experienced anger, whereas the remaining half were asked to describe a normal day in their lives. After finishing this part of the study, participants were told that they would continue with an unrelated study. In this second part they played six times a sequential trust game with the possibility of punishment. In half of these games they encountered a defector and in the other half they encountered a cooperator. Preliminary results by van Dijk, Gallucci, Seip and Rotteveel showed that participants primed with anger punished significantly more than the participants that were not primed with anger, providing further support

for our hypothesis that anger specifically underlies altruistic punishment.

Although there are important differences in studies measuring rejection of offers and those using more direct forms of punishment, together these studies indicate that experienced anger is a more reliable predictor of altruistic punishment than unfairness or defection per se. What seems to matter in altruistic punishment is not the extent to which others defect or how unfair this defection is perceived but the extent to which people experience anger.

Discussion and Future Directions

In the present article we argued that anger constitutes a proximate mechanism for altruistic punishment. In a first attempt to lay down the workings of anger and unfairness in the initiation of altruistic punishment, we argue that the perception of unfairness can cause anger but it is the experience of anger that results in altruistic punishment. Furthermore, we have described some of our recent work that supports the causal role of anger in altruistic punishment.

In a recent study Herrmann and colleagues[12] introduced a new phenomenon in addition to altruistic punishment, namely antisocial punishment (e.g., the sanctioning of people behaving prosocially). This research shows that individuals that have been punished in the past for contributing too little might retaliate against cooperators because they are the ones who are most likely to punish the free-riding low contributors. The authors suggest that at least some people might not accept punishment and therefore seek revenge. Wanting to take revenge is an action tendency closely related to, if not originating from, anger.[30] Therefore, in our view, anger might not only underlie altruistic punishment but also antisocial punishment.

In addition to anger, other emotions could also play a role in altruistic punishment. For example, experienced or anticipated satisfaction following punishment could have an impact on people's willingness to altruistically punish others. In this sense, the push of anger could be complemented possibly by a subsequent pull of satisfaction. In other words, people may be motivated to punish defectors by the satisfaction they derive or expect to derive from imposing punishment upon them. For example, most people seem to feel bad if they observe that free riders are not punished and experience satisfaction if justice is established by punishing defectors.[26,31,32] Moreover, research has shown that punishment activates brain areas related to reward processing[33] and that feelings of (righteous) satisfaction are augmented when the suffering happens to someone who angered them.[31,32]

Emotion is only elicited in response to a human interaction partner and not in response to a computer interaction partner. Although in both situations the offer could be perceived as unfair, it is plausible that the computer partner did not elicit anger and therefore the evaluation of unfairness alone was not sufficient to use altruistic punishment. Sanfey and colleagues[25] showed that unfair offers made by human partners were rejected at a significantly higher rate than those offers made by a computer. The reduced willingness to reject an unfair offer upon applying TMS to the DLPFC only applied to human interaction partners and not to computer interaction partners.[29] The authors argue, based upon previous evidence, that two fairness motives—reciprocity and inequity aversion—are simultaneously activated in the human offer condition whereas only one fairness motive—inequity aversion—is operative in the computer offer condition. Perhaps an unequal offer by a computer is perceived as unfair but does not elicit anger. In interpersonal interaction, you as a person have a certain intention and expect this intention to be reciprocated by your partner. When this intention is not reciprocated, anger is elicited.[34] In support of this, Van 't Wout and colleagues[24] showed that higher skin conductance responses (a proximate for emotion) for unfair compared

to fair offers were only seen for human interaction partners. In investigating the hypothesis that anger underlies altruistic punishment, future research should take intentionality (for both participant and interaction partner) and reciprocity into account.

You expect another to treat you as you would treat him and you perhaps experience more intense emotions when this reciprocity is violated. The highlighted importance of intention might, therefore, also relate to individual differences. People regarded as prosocial (e.g., being concerned with an overall group norm, striving for equality, and trying to maximize joint outcomes) are more likely to reciprocate their partner's action than proselves, who are more motivated to maximize their own outcome.[35] Other personality traits, such as regarding the emotional reaction in response to a violation of reciprocity (unfair behavior), might also influence the decision to use altruistic punishment. Kassinove and colleagues[36] showed that subjects with high levels of trait anger made more competitive/attack responses than did participants with low levels of trait anger; this was supported by our own results where participants primed with anger punished harder.[30] Additional studies should take these individual differences into account because they can provide interesting insights into the prevalence of altruistic punishment among people.

Future research on the role of anger and the perception of unfairness in altruistic punishment could demonstrate whether the effect of anger is unique or whether it is complemented by, for example, (anticipated) satisfaction. In addressing contextual factors we can try to understand when the perception of unfairness also triggers anger. The results of these future investigations will allow us to obtain more details of the underlying mechanism of altruistic punishment.

Acknowledgments

This paper was supported by the MacArthur Foundation Network on Economic Environments and the Evolution of Individual Preferences and Social Norms.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Hamilton, W.D. 1964. Genetical evolution of social behavior I and II. *J. Theor. Biol.* **7:** 1–52.
2. Axelrod, R. & W.D. Hamilton. 1981. The evolution of cooperation. *Science* **211:** 1390–1396.
3. Trivers, R. 1971. The evolution of reciprocal altruism. *Q. Rev. Biol.* **46:** 35–57.
4. Alexander, R.D. 1987. *The Biology of Moral Systems*. Aldine de Gruyter. New York.
5. Nowak, M.A. & K. Sigmund. 1998. The dynamics of indirect reciprocity. *J. Theor. Biol.* **194:** 561–574.
6. Gintis, H., E. Smith & S. Bowles. 2001. Costly signalling and cooperation. *J. Theor. Biol.* **213:** 103–119.
7. Fehr, E. 2002. The nature of human altruism. *Nature* **415:** 269–272.
8. Fehr, E. & U. Fischbacher. 2003. The nature of human altruism. *Nature* **425:** 785–791.
9. Fehr, E. & U. Fischbacher. 2004. Third-party punishment and social norms. *Evol. Hum. Behav.* **25:** 63–87.
10. Fehr, E., & S. Gächter. 2002. Altruistic punishment in humans. *Nature* **415:** 137–140.
11. Henrich, J., R. McElreath, A. Barr, *et al*. 2006. Costly punishment across human societies. *Science* **312:** 1767–1770.
12. Herrmann, B., C. Thöni & S. Gächter. 2008. Antisocial punishment across societies. *Science* **319:** 1362–1367.
13. Boyd, R. & P.J. Richerson. 1992. Punishment allows the evolution of cooperation (or anything else) in sizeable groups. *Ethol. Sociobiol.* **13:** 171–195.
14. Frijda, N.H. 1986. *The emotions*. Cambridge University Press. Cambridge.
15. Frijda, N.H. 2007. *The laws of emotion*. Erlbaum. Mahwah, NJ.
16. Rotteveel, M. & R.H. Phaf. 2007. Mere exposure in reverse: Mood and motion modulate memory bias. *Cognit. Emot.* **21:** 1323–1346.
17. Rowe, G., J.B. Hirsh & A.K. Anderson. 2007. Positive affect increases the breadth of attentional selection. *Proc. Natl. Acad. Sci. Unit. States. Am.* **104:** 383–388.
18. Fredrickson, B.L. 2001. The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. *Am. Psychol.* **56:** 218–226.
19. Harmon-Jones, E. & J.J.B. Allen. 1998. Anger and frontal brain activity: EEG asymmetry consistent

with approach motivation despite negative affective valence. *J. Pers. Soc. Psychol.* **74:** 1310–1316.
20. Rolls, E.T. 2005. *Emotion explained*. Oxford University Press. Oxford.
21. Frank, R.H. 1988. *Passions within reason: The strategic role of the emotions*. Norton. New York.
22. Nesse, R. 1990. Evolutionary explanations of emotions. *Hum. Nat.* **1:** 261–289.
23. Tabibnia, G., A.B. Satpute & M.D. Lieberman. 2008. The sunny side of fairness. Preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). *Psychol. Sci.* **19:** 339–347.
24. Van 't Wout, M., R.S. Kahn, A.G. Sanfey & A. Aleman. 2006. Affective state and decision-making in the ultimatum game. *Exp. Brain. Res.* **169:** 564–568.
25. Sanfey, A.G., J.K. Rilling, J.A. Aronson, *et al*. 2003. The neural basis of economic decision-making in the Ultimatum Game. *Science* **300:** 1755–1758.
26. Singer, T., B. Seymour, J.P. O'Doherty, *et al*. 2006. Empathic neural responses are modulated by the perceived fairness of others. *Nature* **439:** 466–469.
27. Pillutla, M.M. & J.K. Murnighan. 1996. Unfairness, anger, and spite: emotional rejections of ultimatum offers. *Organ. Behav. Hum. Decis. Process.* **68:** 208–224.
28. Ochsner, K.N. & J.J. Gross. 2005. The cognitive control of emotion. *Trends. Cognit. Sci.* **9:** 242–249.
29. Knoch, D., A. Pascual-Leone, K. Meyer, *et al*. 2006. Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science* **314:** 829–832.
30. Frijda, N.H. 1994. The lex talionis: On vengeance. In *Emotions: Essays on emotion theory*. S.H.M. van Goozen, N.E. Van De Poll & J.A. Sergeant, Eds.: 263–289. Erlbaum. Hillside, NJ.
31. Van Dijk, W.W., J.W. Ouwerkerk, S. Goslinga & M. Nieweg. 2005. Deservingness and *Schadenfreude*. *Cognit. Emot.* **19:** 933–939.
32. Van Dijk, W.W., J.W. Ouwerkerk, S. Goslinga, *et al*. 2006. When people fall from grace: Reconsidering the role of envy in *Schadenfreude*. *Emotion* **6:** 156–160.
33. De Quervain, D.J.F., U. Fischbacher, V. Treyer, *et al*. 2004. The neural basis of altruistic punishment. *Science* **305:** 1254–1258.
34. Lamm, H. 1986. Justice considerations in interpersonal conflict. In *Justice in social relations*. H.W. Bierhoff, R.L. Cohen & J. Greenberg, Eds. Plenum Press. New York.
35. De Cremer, D. & P.A.M. van Lange. 2001. Why prosocials exhibit greater cooperation than proselfs: The roles of social responsibility and reciprocity. *Eur. J. Pers.* **15:** 5–18.
36. Kassinove, H., D. Roth, S.G. Owens & J.R. Fuller. 2002. Effects of trait anger and anger expression style on competitive attack responses in a wartime prisoner's dilemma game. *Aggressive Behav.* **28:** 117–125.
37. O'Gorman, R., D.S. Wilson & R.R. Miller. 2005. Altruistic punishment and helping differ in sensitivity to relatedness, friendship, and future interactions. *Evol. Hum. Behav.* **26:** 375–387.