The Evolution of Altruistic Punishment

By Simon Columbus

Abstract: Altruistic punishment has been noted as a force in sustaining cooperation. The evolution of altruistic punishment, however, is hard to explain by natural selection. In this paper, we review the function, causation, and evolution of altruistic punishment. We find that altruistic punishment increases cooperation and, in the long term, total group welfare. Individual welfare, however, is decreased relative to non-punishing cooperation. Negative emotions, in particular anger, can be identified as a proximate cause of punitive behaviour. Cultural group selection has been proposed as an evolutionary pathway for individually non-adaptive altruistic punishment, and is supported by empirical findings.

I. INTRODUCTION

The human species has an outstanding ability to cooperate. To probe its nature, game theory models economic interactions in simple experimental settings. One reliable finding from this line of research is that cooperation breaks down in the absence of enforcement mechanisms (Fehr & Gächter, 2000; 2002). So-called "altruistic" punishment of defectors has been widely found to effectively sustain cooperation (Boyd, Gintis, & Bowles, 2010; Egas & Riedl, 2008; Gächter, Renner, & Sefton, 2008; Gürerk, Irlenbusch, & Rockenbach, 2006; Mathew & Boyd, 2011; O'Gorman, Henrich, & van Vugt, 2009; Ostrom, Walker, & Gardner, 1992) and is believed to be a key to understanding the evolution of cooperation (Botelho et al., 2005; Boyd & Richerson, 1992; Dreber et al., 2008; Eldakar & Wilson, 2008; Fehr & Fischbacher, 2003; Fehr & Rockenbach, 2004; Hauert et al., 2007; Henrich, 2006; Henrich & Boyd, 2001; Nowak & Sigmund, 2005; Ohtsuki, Iwasa, & Nowak, 2009; Panchanathan & Boyd, 2004; Rockenbach & Milinski, 2006; Ule et al., 2009). The evolution of altruistic punishment itself, however, poses an equally fascinating puzzle. This essay will therefore analyse the evolution of altruistic punishment within Tinbergen's (1963) classic ethological framework. The latter suggests that a coherent explanation of behaviour ought to answer four questions, referring to proximal causation, development (ontogeny), evolution (phylogeny), and survival value (function) of observed behaviour. This paper studies the causation, function, and evolution of altruistic punishment. We hypothesize that altruistic punishment increases cooperation and group pay-offs, but decreases individual pay-offs; that negative emotions are a proximate cause of punishment; and that punishment evolved by cultural group selection.

II. BEHAVIOURAL ECONOMICS STUDY THE CONTEMPORARY FUNCTION OF ALTRUISTIC PUNISHMENT

In behavioural economics and related experimental disciplines, altruistic punishment is frequently operationalised in two settings, using anonymous, one-shot Ultimatum games (UG) or more sophisticated, iterated Public Goods (PG) games (for an experimental protocol, see Fehr & Gächter, 2000; 2002). In UG, rejection of highly unfair offers—with a 80:20 or more skewed split—is common (Forsythe et al., 1994), and frequently half or more of such offers are turned down across a wide variety of cultures (Henrich et al., 2006). This form of altruistic punishment co-varies with pro-social behaviour in other games (ibid.; cf. Henrich et al., 2005). Indeed, unfair offers are rare—frequently, less than 20% of players offer an 80:20 or more skewed split (Forsythe et al., 1994; Henrich et al., 2006)—and less common than in Dictator games which do not allow for punishment (e.g. Forsythe et al., 1994), which indicates that most players expect unfair offers to be rejected. In PG games, although it has been shown that even solitary punishers can sustain cooperation (O'Gorman, Henrich, & van Vugt, 2009), punishment is wide-spread, with up to 84% of players punishing at least once over ten rounds (Fehr & Gächter, 2002). The cost/benefit ratio of punishment poses a strong constraint, however, as punishment is only frequent for high (commonly 3:1) ratios (Egas & Riedl, 2008).

Altruistic punishment has a particular role in sustaining cooperation, as evidenced in PG games. In the absence of enforcement mechanisms, cooperation in PG games frequently breaks down after multiple iterations, although initial levels of cooperation, with investments at about 60% of players' endowments, are well above the zero-sum Nash equilibrium (Fehr & Schmidt, 1999; Fehr & Gächter, 2000). In contrast, a punishment option increases average contributions even in stranger (anonymous) treatments, and establishes near-perfect cooperation in partner (non-anonymous) treatments (Fehr & Gächter, 2000). Early studies found that most (74%, Fehr & Gächter, 2002) punishment is targeted at free-riders—i.e. players whose contribution lies below the group average. Cross-cultural studies, however, have since established that in some societies, anti-social punishment—i.e. punishment directed at average or above-average contributors—can be just as frequent as pro-social punishment (Herrmann, Thöni, & Gächter, 2008). High levels of anti-social punishment can effectively lead to the breakdown of cooperation (ibid.). In one study, anti-social punishment was associated with society-wide low adherence to civic norms and weak rule of law (ibid.) and in particular occurred in Arabic-speaking and Southern European cultures (Gächter, Herrmann, & Thöni, 2010).

There is considerable variation in punitive behaviour within subject pools, and studies indicate that players employ a wide variety of strategies. In a PG game without punishment condition, Fischbacher and Gächter classified players as conditional cooperators (55%), free-riders (23%), or *triangle cooperators*, whose contribution increases relative to that of the other players up to a point, then decreases again (12%); a large number of players (10%) could not be classified. A study by Ule et al. (2009), which allowed for both punishing and rewarding players, established as many as nine strategic categories along the axes of self/other-regarding and discriminate/indiscriminate behaviour; their frequency varied according to whether punishment was effective or merely symbolic. Notably, with reward also being an option, punishment was much less frequent (13.2% vs. 56.6% in the harmful punishment condition). Socio-economic factors generally do not seem to explain differences in punitive behaviour (Henrich et al., 2006; although Egas & Riedl (2008) indicate older men punish more heavily).

Players' individual strategic choice has significant impacts on pay-offs, with punishers earning lower individual pay-offs than other players (Dreber et al., 2008; Ule et al., 2009). Indeed, despite its cooperation-enhancing effect, punishment seems to decrease total welfare (Fehr & Gächter, 2002; Gürerk, Irlenbusch, & Rockenbach, 2008; Herrmann, Thöni, & Gächter, 2008; Egas & Riedl, 2008; Dreber et al., 2008); however, Gächter, Renner, and Sefton (2008) have shown that this effect reverses over many iterations as deviations, and subsequent punishment, become less frequent.

III. PSYCHOLOGICAL MECHANISMS ARE PROXIMAL CAUSES ALTRUISTIC PUNISHMENT

The frequency of cooperative behaviour in economic games, including altruistic punishment, has informed models explicitly meant to replace the rationally self-interested *homo economicus* in mainstream economics. These models frequently maintain the rationality assumption, but posit a variety of social preferences. Fehr and Schmidt (1999) have proposed a utility curve that incorporates inequality aversion, i.e. a preference for equality among participants in a game. Indeed, subjects punished high earners in an anonymous give/take scenario without cooperation, and rewarded low earners (Dawes et al., 2007), which indicates that inequality aversion, rather than enhancement of cooperation, motivates punishment.

Negative emotional reactions have early been suggested as a mechanism for social preferences (Fehr & Gächter, 2002). Indeed, Dawes et al. (2007) find that players report being angry at high earners, which is also supported by skin conductance measures (van't Wout, Kahn, & Sanfey, 2006; Seip, van Dijk, & Rotteweel, 2009). In a study by Seip, van Dijk, and Rotteweel (2009), reported anger fully mediated the effect of perceived unfairness on punishment intensity, indicating a primacy of emotion over deliberate norm enforcement. Priming subjects with anger, too, increased punishment (ibid.). In line with Seip, van Dijk, and Rotteweel's (2009) finding that anger, rather than perceived unfairness, predicts punishment, Knoch et al. (2006) showed that disrupting the right prefrontal cortex increased acceptance rates for unfair offers in the UG, but did not impact fairness ratings, i.e. dissociated fairness judgements and punishment. Jensen (2010) interprets previous findings to indicate that humans are, at least in part, motivated by spite, i.e. have negative social preferences. Concurrently, Houser and Xiao (2010) provide some support for the existence of inequality-seeking punishment.

Neurological measures appear to indicate a conflict between cognitive and emotional responses to unfairness underlying punishment. Among the brain areas frequently implicated in altruistic punishment are regions of the prefrontal cortex (PFC) such as the dorsolateral PFC (Strobel et al., 2011) and ventromedial PFC and medial orbitofrontal cortex (de Quervain et al., 2004), which play a role in cognitive control, as well as the insula, which has been proposed to be involved in the representation of negative emotional states (Sanfey et al., 2003; Strobel et al., 2011).

In a variety of studies using different protocols and methods, virtual lesioning of the right dlPFC decreased punishment (Knoch et al., 2006; 2008), and lower baseline activity in the rPFC predicted decreased punishment (Knoch et al., 2010). However, damage to the vmPFC was associated with increased punishment (Koenigs & Tranel, 2007). Intriguingly, Sanfey et al. (2003) reported that punishment followed lower activation in the dlPFC than in the insula, and vice-versa. They interpreted this finding as evidence for competition between cognitive processes underlying maximisation of monetary reward and the emotional anger response. However, these findings stand in stark contrast to most later studies showing a decrease in punishment following suppression of dlPFC activation.

Strobel et al. (2011) have suggested that the insula provides a *bias signal* to the striatum, which possibly represents anticipated satisfaction from punishing (de Quervain et al., 2004; Fehr, Fischbacher, & Kosfeld, 2005; Seymour, Singer, & Dolan, 2007; Strobel et al., 2011). The behavioural conflict that arises from contradictory emotional and cognitive motivations might be monitored and signalled in the anterior cingulate cortex (ACC, Sanfey et al., 2003; Strobel et al., 2011), and ventromedial and orbitomedial PFC have been proposed for the goal-directed integration of these signals (de Quervain et al., 2004; Koenigs & Tranel, 2007; Seymour, Singer, & Dolan, 2007). These findings indicate that altruistic punishment involves both cognitive factors arising from expected monetary reward and emotional motivations stemming from the expected satisfaction derived from punishing. However, current evidence on neural substrata underlying altruistic punishment is sparse and should be considered preliminary.

IV. THE EVOLUTION OF ALTRUISTIC PUNISHMENT CAN BE EXPLAINED BY CULTURAL GROUP SELECTION

Findings that indicate social preferences are difficult to accommodate with standard gene- or individual-based evolutionary theory. The behaviour observed in economic games is true altruism in the sense that players who punish, increasing total welfare for the group in the long term (Gächter, Renner, & Sefton, 2008), incur relative losses for themselves. This has been taken by some as evidence for strong reciprocity (Gintis, 2000; Gintis et al., 2003). As Gintis defines it, "a strong reciprocator is predisposed to cooperate with others and punish non-cooperators, even when this behavior cannot be justified in terms of self-interest, extended kinship, or reciprocal altruism."

To explain the evolution of strong reciprocity, models of cultural group selection have been developed (e.g. Boyd & Richerson, 2009). It is presumed that culture allows for local adaptation at a rate faster than evolution by natural selection, which - if mixing between groups is infrequent - gives rise to heritable variation across groups. Then, this stable variation among groups leads to group selection, which means not necessarily the extinction of groups, but of cultural practices; in the end, cooperative (pro-social) practices prevail. Three mechanisms by which culture can spread are intergroup competition, human propensity to imitate the successful, and selective migration. By these means, cultural evolution would give rise to cooperative groups and create an environment in which, in turn, the evolution of traits that give an advantage in a social setting would be favoured. Boyd et al. (2003) specifically argue that cultural group selection can give rise to altruistic punishment. They show that if defectors are rare in a population, punishers have only a weak pay-off disadvantage compared to non-punishing cooperators. If cooperative groups have a higher survival rate than non-cooperative groups in which punishers are absent, group selection would allow for punishment to spread.

Several empirical findings lend support to the propositions made by Boyd, Richerson, and their colleagues. The pay-off disadvantage of punishers indeed shrinks in the long term as defection and punishment become increasingly rare (Gächter, Renner, & Sefton, 2008). Gürerk, Irlenbusch, and Rockenbach (2006) show that sanctioning institutions have a competitive advantage, providing an experimental case of cultural group selection. Given the choice between two PG games with and without punishment condition, most participants initially choose the latter; but switch groups as it becomes apparent that cooperation breaks down without punishment. These two findings would seem to support two basic tenets of cultural group selection; however, it has to be noted that current experimental findings cannot reproduce human interactions before the evolution of altruistic punishment, and are thus of limited use when providing evidence for such evolutionary theories. One of the most important implications of Boyd et al.'s theory is that groups will have mixed populations, as neither all-cooperators nor all-punishers are evolutionary stable. Indeed, empirical findings (e.g. Ule et al., 2009) provide evidence that punishers make up only a fraction of any population. An interesting, related finding is that a subject's variant of the MAO-A gene dimorphism, which has been linked to aggression, predicts punishment (McDermott et al., 2009). This provides further support for diverse populations rather than a universality of punishment.

A particular challenge to Boyd et al.'s (2003; 2009) explanation of the evolution of altruistic punishment remains the fact that this behaviour creates a second-order collective action problem. It is not clear why punishers would not be replaced by cooperators-only who, with respect to punishment, are second-order free-riders. Reciprocal altruism, as proposed by Panchanathan and Boyd (2004), ultimately does not provide a solution. Lastly, the role of framing (Hagen & Hammerstein, 2006) remains an open question. The important role played by culture, emotions, and implicit as well as explicit game features, however, suggests that this role is not to be neglected.

VI. REFERENCES

Botelho, A., Harrison, G. W., Pinto, L., & Rutström, E. E. (2005). Social norms and social choice. Working Paper, Universidade do Minho.

Boyd, R., Gintis, H., & Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science, 328*(5978), 617-20.

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences of the United States of America, 100*(6), 3531-5.

Boyd, R. & Richerson, P. J. (1992). Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizable Groups. *Ethology and Sociobiology, 13*, 171-95.

Boyd, R. & Richerson, P. J. (2009). Culture and the evolution of human cooperation. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 364*(1533), 3281-8.

Buckholtz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., & Marois, R. (2008). The neural correlates of third-party punishment. *Neuron, 60*(5), 930-40.

Dawes, C. T., Fowler, J. H., Johnson, T, McElreath, R., & Smirnov, O. (2007). Egalitarian motives in humans. *Nature, 446*(7137), 794-6.

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*(5688), 1254-8.

Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don't punish. *Nature, 452*(7185), 348-51.

Eldakar, O. T. & Wilson, D. S. (2008). Selfishness as second-order altruism. *Proceedings of the National Academy of Sciences of the United States of America, 105*(19), 6982-6.

Elliott, C. S. & Hayward, D. M. (1998). The expanding definition of framing and its particular impact on economic experimentation. *Journal of Socio-Economics, 27*(2), 229-43.

Egas, M. & Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society. Series B, Biological Sciences, 275*(1637), 871-8.

Fehr, E. & Fischbacher, U. (2003). The nature of human altruism. *Nature, 425*(6960), 785-91.

Fehr, E., Fischbacher, U., & Kosfeld, M. (2005). Neuroeconomic Foundations of Trust and Social Preferences: Initial Evidence. *The American Economic Review, 95*(2), 346-51.

Fehr, E. & Gächter, S. (2000). Cooperation and Punishment in Public Goods Experiments. *The American Economic Review, 90*(4), 980-994.

Fehr, E. & Gächter, S. (2002). Altruistic Punishment in Humans. *Nature, 415*(6868), 137-40.

Fehr, E. & Schmidt, K. (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics, 114*(3), 817-68.

Gächter, S., Herrmann, B., & Thöni, C. (2010). Culture and Cooperation. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 365*(1553), 2651-61.

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

Gürerk, O., Irlenbusch, B., & Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science, 312*(5770), 108-11.

Hagen, E. H. & Hammerstein, P. (2006). Game theory and human evolution: a critique of some recent interpretations of experimental games. *Theoretical Population Biology, 69*(3), 339-48.

Hauert, C., Traulsen, A., Brandt, H., Nowak, M. A., & Sigmund, K. (2007). Via freedom to coercion: the emergence of costly punishment. *Science, 316*(5833), 1905-7.

Henrich, J. (2006). Cooperation, punishment, and the evolution of human institutions. *Science, 312*(5770), 60-1.

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*(1), 79-89.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F. W., Tracer, D., & Ziker, J. (2006). Costly punishment across human societies. *Science, 312*(5781), 1767-70.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial Punishment Across Societies. *Science, 319*(5868), 1362-7.

Houser, D. & Xiao, E. (2010). Inequality-seeking Punishment. *Economics Letters, 109*(1), 20-23.

Jensen, K. (2010). Punishment and spite, the dark side of cooperation. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 365*(1553), 2635-50.

Knoch, D., Gianotti, L. R. R., Baumgartner, T., & Fehr, E. (2010). A neural marker of costly punishment behavior. *Psychological Science, 21*(3), 337-42.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science, 314*(5800), 829-32.

Koenigs, M. & Tranel, D. (2007). Irrational economic decision-making after ventromedial prefrontal damage: evidence from the Ultimatum Game. *The Journal of Neuroscience, 27*(4), 951-6.

Mathew, S. & Boyd, R. (2011). Punishment sustains large-scale cooperation in prestate warfare. *Proceedings of the National Academy of Sciences of the United States of America, 108*(28), 11375-80.

McDermott, R., Tingley, D., Cowden, J., Frazzetto, G., & Johnson, D. D. P. (2009). Monoamine oxidase A gene (MAOA) predicts behavioral aggression following provocation. *Proceedings of the National Academy of Sciences of the United States of America, 106*(7), 2118-23.

Nowak, M. A. & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature, 437*(7063), 1291-8.

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. Series B, Biological Sciences, 276*(1655), 323-9.

Ohtsuki, H., Iwasa, Y., & Nowak, M. A. (2009). Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. *Nature, 457*(7225), 79-82.

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants With and Without a Sword: Self-Governance is Possible. *The American Political Science Review, 86*(2), 404-417.

Panchanathan, K. & Boyd, R. (2004). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature, 432*(7016), 499-502.

Rockenbach, B. & Milinski, M. (2006). The efficient interaction of indirect reciprocity and costly punishment. *Nature, 444*(7120), 718-23.

Sanfey, A. G., Rilling, J. K., Aronson, J. A, Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science, 300*(5626), 1755-8.

Seip, E. C., van Dijk, W. W., & Rotteveel, M. (2009). On hotheads and Dirty Harries: the primacy of anger in altruistic punishment. *Annals of the New York Academy of Sciences, 1167*, 190-6.

Seymour, B., Singer, T., & Dolan, R. (2007). The Neurobiology of Punishment. *Nature reviews. Neuroscience, 8*(4), 300-11.

Strobel, A., Zimmermann, J., Schmitz, A., Reuter, M., Lis, S., Windmann, S., & Kirsch, P. (2011). Beyond revenge: neural and genetic bases of altruistic punishment. *NeuroImage, 54*(1), 671-80.

Tinbergen, N. (1963). On aims and methods of Ethology. *Zeitschrift für Tierpsychologie, 20*(4), 410-33.

Tversky, A. & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453-8.

Ule, A., Schram, A., Riedl, A., & Cason, T. N. (2009). Indirect punishment and generosity toward strangers. *Science, 326*(5960), 1701-4.

van't Wout, M., Kahn, R. S., Sanfey, A. G., Aleman, A. (2006). Affective state and decision-making in the Ultimatum Game. *Experimental Brain Research, 169*(4), 564-8.