THE NEURAL BASIS OF ALTRUISTIC PUNISHMENT

By Simon Jonas Hadlich

Abstract: Altruistic punishment can allow human groups to sustain cooperation. This paper reviews the research on the neural basis of punitive behaviour. Altruistic punishment involves cognitive brain regions such as the dorsolateral prefrontal cortex as well as areas representing negative emotions, such as the insula. The conflict between cognitive and emotional inputs is likely monitored, signalled, and integrated in the anterior cingulate cortex, as well as the ventromedial and orbitomedial prefrontal cortex. Altruistic punishment is also associated with reward-related activation in the striatum. These findings have implications for evolutionary theory, the foundations of microeconomics, and institutional design.

Keywords: altruism; punishment; behavioural neuroscience; cooperation.

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). Consequently, the human tendency to punish free-riders even at personal cost has important implications for cooperative endeavours today. To enlighten the mechanisms that underlie punishment, this paper will link neural and behavioural research.

**Measuring Punishment: Methodological Considerations**

Altruistic punishment is usually defined as punishing perpetrators of anti-social or norm-digressing behaviour at a personal cost. This concept is frequently operationalised in one of two experimental set-ups. The first set-up is based on the Ultimatum game, which is played in anonymous one-shot interactions between two players. Player A receives a certain endowment. They can then decide to split that endowment with Player B. Subsequently, Player B can decide to either accept or reject the split. In case of acceptance, both players receive their respective share; in case of rejection, both do not receive anything. In anonymous one-shot interactions, Player B’s rejection is interpreted as altruistic punishment: Player B accepts a personal cost to themselves to rid Player A of any benefits, but any learning effect of A’s side will not have implications for B. Analysed in terms of game theory, such behaviour is irrational: any non-zero offer dominates rejection, and should thus be accepted. This also implies that no rationally self-interested Player A should offer more than a marginal non-zero sum, anticipating that an equally rational Player B would accept any such offer. This method, using Ultimatum Games, has been the chief operationalisation of altruistic punishment in neuroscience studies. It has some major disadvantages, however, including that it does not allow for variations in punishment magnitude. It might also be that subjects do not truly believe, or comprehend, one-shot conditions.

The second, more sophisticated - and arguably preferable - operationalisation of altruistic punishment is a variation of the Public Goods Game, which can be played over multiple iterations. In this experiment, a group of players - commonly between 3 and 6 - each receive a certain endowment. Each can then decide to invest a share of the sum in a "common project", a pot in which money is pooled. Players receive any money they do
not invest in the common project. Money paid into the pot is multiplied by a certain factor $k$ ($1 < k < n$, commonly 1.4 - 1.6) and paid out in equal shares to all players. Here, game theory would predict no invest into the pot from any players. Since each player can maximize their returns by investing nothing while others pay some money into the pot, an investment of 0 strictly dominates all other choices. In a punishment condition, after all players have received their share from the pot, each can invest a set sum of money to deduct some amount from any other player's account. Commonly, the cost of punishment is less than its effect (ratios of 1:3 are frequently found in the literature). This does not change the game theory prediction for the first part of the game, as it indeed creates a second order collective action problem: again, free-riding on other players' expenses on punishment is dominant, meaning that no player should rationally spend on punishment. Public Goods Games have largely been used in economic experiments, in particular the work of Fehr and his colleagues (Fehr & Gächter, 2000; 2002).

The actual outcomes of Ultimatum and Public Goods Games if played by experimental subjects strongly contradict game theoretic predictions about rationally self-interested behaviour. In Ultimatum Games, frequently more than half of all players offer a 50:50 split, and fewer than 20% offer 80:20 or an even more skewed split (Forsythe et al., 1994). While there is some variation across cultures, there is overwhelming evidence that non-zero Ultimatum Game offers are universal (Henrich et al., 2006). Rejection rates are subject to higher variation, but are frequently as high as 50% for offers less equal than an 80:20 split (ibid.). In Public Goods Games, despite the unpromising game theory prediction, initial investments are as high as 60% on average; however, they decrease swiftly towards zero over multiple iterations. If a punishment condition is introduced, average contributions increase up to near-perfect cooperation in partner (non-anonymous) treatments, and are high even in stranger treatments (Fehr & Gächter, 2000). Punishment is wide-spread, with up to 84% of players punishing at least once over ten rounds (Fehr & Gächter, 2002), and most (74%) are targeted at players who contributed less than the average (ibid.). While this cooperation-enhancing effect of punishment has been shown across various societies (Herrmann, Thöni, & Gächter, 2008), punitive behaviour differs widely across societies (ibid.). In some locations - in particular the Middle East and Southern Europe - punishment directed at
average and high contributors to a similar degree as at defectors can lead to a breakdown in cooperation (Gächter, Herrmann, & Thöni, 2010).

**Neural Correlates of Punitive Behaviour**

The dorsolateral prefrontal cortex (DLPFC) is activated during the decision to punish. In particular, this brain region is active during the assessment of guilt in another player (Buckholtz et al., 2008). Activation is strongest for unfair (as opposed to fair) offers, and when subjects to be punished are responsible for their decision (ibid.; Sanfey et al., 2003). There is some evidence that activation in the DLPFC is directly and positively related to the incidence of rejection of unfair offers in the Ultimatum Game. A disruption of the right (but not left) DLPFC is associated with a lower rate of rejection of unfair offers (Knoch et al., 2006), and right lateral PFC baseline hypoactivation predicts heightened acceptance rates of unfair offers (Knoch et al., 2010). Knoch and her colleagues thus conjecture that baseline activation in the right DLPFC is a neural marker for altruistic punishment (ibid.). A more intricate regulatory mechanism might exist, however, as "unfair offers that are subsequently rejected have greater anterior insula than DLPFC activation, whereas accepted offers exhibit greater DLPFC than anterior insula" activation (Sanfey et al., 2003, p. 1757). Thus the relationship observed by Knoch et al. (2006; 2010) would be modulated by activity in the insula. Prior research strongly indicates that the DLPFC is involved in the implementation of cognitive control and the mediation of goal-directed action across time; consequently, it might take a regulatory role in altruistic punishment (Seymour, Singer, & Dolan, 2007; Strobel et al., 2011).

Unfair offers elicit greater activation in the bilateral insula than do fair offers (Sanfey et al., 2003), and activation is stronger if unfair offers were made by a human player rather than a computer (ibid.). There is also a link between insular activation and rejection of offers in the Ultimatum Game. Subjects with higher anterior insula activation reject more unfair offers (ibid.), and activation in this region is higher for offers that are later rejected (ibid.; Strobel et al., 2011). Strobel et al. (2011) argue that the insula represents negative emotional states arising from norm infractions, sending out a "bias signal". As Sanfey et al. (2003) note, insular activity seems to counter DLPFC activation, such that the stronger one of the two determines behavioural outcomes. For
unfair offers that are later rejected, insula activation is greater than DLPFC activation (ibid.), which could consequently be interpreted as the insula's bias signal overriding the incentive of monetary reward.

In the anterior cingulate cortex, greater activation is elicited by unfair (as compared to fair) offers (Sanfey et al., 2003), and activation positively correlates with punishment of unfair offers (Strobel et al., 2011). Prior research indicates that this region is involved in the detection of cognitive conflict. It monitors and signals behavioural conflict that arises from a possible bias signal provided by the insula and signals of monetary reward, i.e. the conflicting cognitive and emotional motivations that exist in the Ultimatum Game (Sanfey et al., 2003; Strobel et al., 2011).

The ventromedial prefrontal cortex (VMPFC) is activated in situations where punishment poses a decision problem. De Quervain et al. (2004) find that VMPFC activation is higher when punishment is costly to the punisher (as opposed to "free" punishment), i.e. when the value of punishment needs to be weighed against its cost. Prior research indicates that the VMPFC is involved in the goal-directed integration of multiple cognitive operations, which in this case would be monetary and punitive value (ibid.). It is also associated with emotional control. Lesions in the VMPFC are linked to impaired modulation of emotions, in particular social frustration, which includes "exaggerated anger, irritability, [and] emotional outbursts" (Koenigs & Tranel, 2007, p. 954). Indeed, subjects with VMPFC lesions show strongly increased rejection rates of unfair (but not fair) offers in the Ultimatum Game (as compared to healthy and subjects with non-VMPFC lesions). Another punishment-related finding shows that the VMPFC is activated during a decision on the magnitude of punishment, even though the latter is not costly (Buckholtz et al., 2008). Taken together, these findings indicate that the VMPFC plays a role in weighing emotional responses to unfair offers against the monetary cost of rejection. At least one study also ascribes a similar role to the medial orbitofrontal cortex (medial OFC), which might represent retributive value (de Quervain et al., 2004; Seymour, Singer, & Dolan, 2007).

The striatum experiences stronger activation during punishment, and there is a positive correlation between striatal activation and more severe punishment (de Quervain et al., 2004). The nucleus caudatus (NCd) has also been shown to be more strongly activated in effective punishment conditions (as opposed to merely symbolic
punishment) (de Quervain et al., 2004; Strobel et al., 2011). The dorsal striatum might control the execution of punishment acts and might be involved in learning punitive behaviour, whereas the nucleus accumbens (NAc) in the ventral striatum might represent retributive value (Seymour, Singer, & Dolan, 2007). The dorsal striatum is understood to be the reward centre of the brain, with a major role in integrating reward-related input signals (Strobel et al., 2011). De Quervain et al. (2004) argue that "caudate activation reflects anticipated satisfaction from punishing defectors" (p. 1258). If punishment was indeed rewarding in this way, subjects with stronger caudate activation would feel compelled to exert more severe punishment, as observed by de Quervain and his colleagues. Fehr, Fischbacher, and Kosfeld (2005) argue that, cast in economic terms, striatal activation is "nonpecuniary utility" derived from punishing norm-defiant behaviour. This utility can directly be weighed against the utility potentially derived from the money gained by accepting an (unfair) offer.

A range of further brain regions have been implicated to be involved in punishment, including anterior intraparietal sulcus (aIPS) and temporo-parietal junction (TPJ) in the decision whether to punish, and posterior cingulate, temporal pole, dorsomedial prefrontal cortex (DMPFC), and inferior frontal gyrus in decisions on the magnitude of punishment (Buckholtz et al., 2008). Studies also show that there are no neural sex differences (Strobel et al.) as well as little difference between first-person and third-party punishment conditions (ibid.).

Some specific neurotransmitters have been implicated in punitive behaviour. Depletion of serotonin, a neurotransmitter whose incidence is generally positively associated with self-control, correlates with increased punishment (Crockett et al., 2010). In contrast, dopamine is positively linked to punishment. Bearers of a COMT gene variant linked to elevated synaptic dopamine availability punish more often (Strobel et al., 2011). These subjects experience greater punishment-related activation in the nucleus accumbens, which might provide a mechanism for this effect (ibid.). In subjects with Parkinson's disease who exhibit impulsive compulsive behaviour, punishment becomes more frequent under the effect of dopaminergic medication (Djamshidian et al., 2011). This medication increases impulsive choice in this group (ibid.), which itself has been shown to correlate with higher rejection rates in the Ultimatum Game (Crockett et al., 2010).
The link between the monoamine oxidase A (MAO-A) gene dimorphism and aggression is among the best understood gene-behaviour relationships. The effect on punishment was studied by McDermott et al. (2009) using a "power-to-take" game in which an anonymous Player B would take away some amount from Player A's earned money, whom A could then punish by administering a painful dose of hot sauce. For large (20:80) takings, i.e. "if significantly provoked", subjects with a low-functioning MAO-A variant punish more often and more intensely than subjects with a high-functioning variant. McDermott and colleagues argue that this is an effect of psychological hypersensitivity to social rejection.

**Modelling the Neural Processes Underlying Punishment**

Neuroimaging studies indicate that lower DLPFC activity is linked to decreased punishment (Knoch, 2006; 2010); however, Sanfey et al. (2003) and Strobel et al. (2011) find that punishment follows greater activity in the DLPFC than in the insula, and vice-versa. While it seems plausible that the DLPFC implements cognitive control, these conflicting findings obscure the exact mechanism. Furthermore, the insula has also been proposed as representing (negative) emotional states and providing a respective bias signal (Strobel et al., 2011). The anterior cingulate cortex monitors and signals behavioural conflict arising from contradictory emotional and cognitive motivations (Sanfey et al., 2003; Strobel et al., 2011). The ventromedial and orbitomedial PFC are involved where punishment poses a decision problem; they are likely involved in the goal-directed integration of cognitive and emotional signals (de Quervain, 2004; Koenigs & Tranel, 2007; Seymour, Singer, & Dolan, 2007). Striatal activation is associated with punishment severity, and as a reward centre in the brain this region possibly represents anticipated satisfaction from punishing (de Quervain et al., 2004; Fehr, Fischbacher, & Kosfeld, 2005; Seymour, Singer, & Dolan, 2007; Strobel et al., 2011). Indicating the role of neurotransmitters, increased synaptic dopamine availability has been linked to greater striatal activation (Strobel et al., 2011).

A range of models have been proposed to link individual findings from neuroimaging stories into a coherent explanation of the neural processes underlying altruistic punishment. An early model by Sanfey et al. (2003) is based on a contrast
between emotional motivations represented in the insula and cognitive ones represented in the DLPFC. For unfair offers, stronger insular than DLPFC activation is associated with later rejection, whereas stronger DLPFC than insular activation is linked to acceptance. Integrating these findings into a more complex framework, Strobel et al. (2011) link activity in the insula and DLPFC to activation of the ACC and striatum:

"[I]nsular representations of negative emotional states due to norm violations would provide a bias signal, which interferes with signals of immediate individual financial reward if no punishment is exerted. The resulting behavioural conflict, monitored and signalled by the ACC, would result in DLPFC-mediated implementation of cognitive control, which would impact on striatal integration of the input signals in favour of the decision to punish, given that more future reward would be anticipated following such behaviour due to learned contingencies between norm-conform or norm-enforcing behaviour and social reward."

Strobel et al.’s model conspicuously omits the VMPFC, which has been implicated in the integration of conflicting inputs. It also conflicts with de Quervain et al.’s (2004) position that striatal activation represents anticipated reward, rather adopting Seymour, Singer, and Dolan’s (2007) perspective that the striatum controls the execution of punitive behaviour. It thus seems necessary to update Strobel et al.’s model accordingly; however, the conflicting evidence regarding the role of the DLPFC - whose involvement itself is perhaps the best documented of all regions - makes this difficult. Further research will be necessary to clarify this question before a comprehensive model can be constructed. Future models should also integrate the role of specific neurotransmitters, such as dopamine and serotonin, which appear to have contrary effects.

**Discussion**

The beginning understanding of the neural basis of altruistic punishment has implications that reach far beyond neuroscience. Punishment has been hypothesised as a key mechanism in the evolution of cooperation, and research into its neural basis
provides input for population models. One of the difficulties that arise in the course is the definition of altruism. In an evolutionary sense, true altruism is defined as behaviour that benefits non-kin at the expense of an individual’s own reproductive fitness. In psychological terms, however, a pleasurable - if fitness-decreasing - behaviour would hardly be defined as altruism. The finding that the anticipation of punishing defectors activates the brain's reward centre in the striatum thus highlights the divergence of biological and psychological altruism.

The involvement of the striatum in punishment also challenges economic theory. As Fehr, Fischbacher, and Kosfeld (2005) cast it, striatal activation can be understood as "nonpecuniar utility". This expresses the understanding that humans value not only (monetary) outcomes, but also processes that lead up to them; an idea that is alien to mainstream economic theory. Together with other, similar preferences - e.g. against inequality - the human propensity to altruistically punish defectors forms the cluster of social preferences (ibid.). These are part of a major update of the assumptions that underlie mainstream economic theory, which aims to replace rational self-interest with the assumption of rationality and social preferences (Fehr & Fischbacher, 2003; Fehr & Rockenbach, 2004; Gintis et al., 2005).

Social preferences in economic games have been interpreted as evidence for strong reciprocity in humans, i.e. cooperative behaviour "that cannot be justified in terms of self-interest, extended kinship, or reciprocal altruism" (Gintis, 2000, p. 169). Since the evolution of strong reciprocity by natural selection is unlikely, cultural group selection has been put forward as an alternative evolutionary mechanism (e.g. Boyd & Richerson, 2009; Boyd et al., 2003). However, Hagen and Hammerstein (2006) offer a powerful critique of this argument, claiming that it is implicitly based on a cognitive model that takes preferences as essential. Classical evolutionary perspectives, however, suggest that strategies, rather than preferences, are essential, which implies higher adaptiveness to environmental conditions. Hagen and Hammerstein therefore suggest further research into the frames applied in punitive decision-making, a research program that is still outstanding.

Finally, understanding altruistic punishment has implications for institutional design. It is perhaps not a mere coincidence that the first Public Goods Game experiment to allow for altruistic punishment was conducted by Ostrom, Walker, and
Gardner (1992) to substantiate game theoretic analyses about the viability of common property regimes. The human propensity to punish free-riders, by whichever way it evolved, provides an avenue for solving collective action problems that mainstream economic theory long ignored or dismissed with reference to the supposed "tragedy of the commons" (Hardin, 1968). In fact, punishment can support cooperation, and sanctioning institutions are fundamental to collective action (Ostrom, 1990). A better understanding of its neural basis will inform the design of institutions that enable communities to govern themselves.

References


