Evolution and Human Behavior 36 (2015) 323–330

Contents lists available at ScienceDirect

Evolution and Human Behavior

journal homepage: www.ehbonline.org

Original Article

Human punishment is motivated by both a desire for revenge and a desire for equality

Jonathan E. Bone [a], Nichola J. Raihani [b,*]

[a] CoMPLEX, University College London
[b] Department of Genetics, Evolution and Environment, University College London

ARTICLE INFO

Article history:
Initial receipt 6 February 2014
Final revision received 17 February 2015

Keywords:
Cooperation
Punishment
Inequity aversion
Altruism
Economic game
Revenge
Inequality aversion

ABSTRACT

Humans willingly pay a cost to punish defecting partners in experimental games. However, the psychological motives underpinning punishment are unclear. Punishment could stem from the desire to reciprocally harm a cheat (i.e. revenge) which is arguably indicative of a deterrent function. Alternatively, punishment could be motivated by the desire to redress the balance between punisher and cheat. Such a desire for equality might be more indicative of a fitness-leveling function. We used a two player experimental game to disentangle these two possibilities. In this game, one player could choose to steal $0.20 from their partner. Depending on the treatment, players interacting with a stealing partner experienced either advantageous inequality, equal outcomes or disadvantageous inequality. Players could punish stealing partners, but some players had access to effective punishment (1:3 fee to fine) whereas others could only use ineffective punishment (1:1). Players who had access to effective punishment could reduce disadvantageous inequality by tailoring their investment in punishment whereas ineffective punishment did not change the relative payoffs of the individuals in the game but could be used to exact revenge. Players punished regardless of whether stealing created outcome inequality or whether punishment was ineffective at removing payoff differentials, suggesting that punishment was at least partly motivated by the desire to inflict reciprocal harm. However, in the effective punishment condition, players' tendency to punish increased if stealing resulted in disadvantageous inequality and, when possible, punishers tailored their investment in punishment to create equal outcomes. Together these findings suggest that punishment is motivated by both a desire for revenge and a desire for equality. The implications of these findings are discussed.

1. Introduction

Punishment typically involves paying a cost to harm individuals who harm or withhold benefits from the punisher (hereafter 'defectors', Clutton-Brock & Parker, 1995; Raihani, Thornton, & Bshary, 2012; but see Irwin & Horne, 2013; Sylwester, Herrmann, & Bryson, 2013 for punishment aimed at helpful or cooperative individuals). Since punishment is costly to administer, both in terms of executing the punishment itself and in terms of the possibility of provoking retaliation from the target (Dreber, Rand, Fudenberg, & Nowak, 2008; Herrmann, Thöni, & Gächter, 2008; Janssen & Bushman, 2008; Nikiforakis, 2008), considerable effort has been expended in trying to understand the evolved function of punitive sentiments (McCullough, Kurzban, & Tabak, 2013; Price, Cosmides, & Tooby, 2002). Specifically, it has been argued that understanding the contexts that reliably motivate punishment can provide key insights into its likely evolved function (Price et al., 2002). Two broad functional explanations have been proposed. First, it has been suggested that punitive sentiment could confer a selective advantage if punishment deters targets (or bystanders) from harming the punisher in future interactions (e.g. dos Santos, Rankin, & Wedekind, 2011; Hilbe & Sigmund, 2010; McCullough et al., 2013). Under this hypothesis (hereafter the 'revenge' hypothesis), individuals should be motivated to reciprocally harm individuals that intentionally harm them, even if punishment cannot immediately equalize the payoffs between the defector and the punisher (Falk, Fehr, & Fischbacher, 2005). However, evidence that punitive sentiments are sensitive to the risk of suffering a fitness disadvantage relative to defectors (Dawes, Fowler, Johnson, McElreath, & Smirnov, 2007; Raihani & McAuliffe, 2012a) suggests an alternative explanation: that punishment primarily serves a fitness-leveling function, by reducing payoff differentials between defectors and punishers (Price et al., 2002; see also Carlsmith, Darley, & Robinson, 2002). Under this fitness-leveling hypothesis, punishers are expected to be motivated primarily by the desire to equalize payoffs, and any deterrent function of punishment would arise as a by-

* Corresponding author. Department of Genetics, Evolution and Environment, University College London, London, WC1E 6BT, United Kingdom.
E-mail address: nicholaraihani@gmail.com (N.J. Raihani).

http://dx.doi.org/10.1016/j.evolhumbehav.2015.02.002
1090-5138/

product. Here, we present an experiment to test whether punitive sentiment can best be explained in terms of desire for revenge or in terms of a desire to equalize payoffs in social interactions.

Interacting with a defector often reduces cooperators' payoffs and creates unequal outcomes. It can therefore be difficult to establish whether punishment of defectors is motivated by the disutility associated with receiving lower payoffs than a defector ('disadvantageous inequality aversion', (Fehr & Schmidt, 1999) or simply a desire for revenge (Raihani & McAuliffe, 2012b). A recent study attempted to disentangle these two possible motivations by asking whether, in the absence of disadvantageous inequality, experiencing losses was sufficient to motivate punishment (Raihani & McAuliffe, 2012b). Raihani and McAuliffe (2012b) found that defection, in the form of stealing money from the victim, did not motivate punishment when stealing resulted in equal outcomes or advantageous inequality for the victim. However, stealing did motivate punishment when it resulted in disadvantageous inequality for the victim (Raihani & McAuliffe, 2012b). These findings raise the possibility that individuals use punishment to restore equality in social interactions. However, the alternative possibility, that punishment is simply related to the disutility associated with experiencing disadvantageous inequality and is not tailored to achieve equal outcomes, could not be ruled out because players in this game were not allowed to tailor their investment in punishment.

Alternative studies have also suggested that investment in punishment is aimed at producing equal outcomes in social interactions. For example, in (Dawes et al., 2007) individuals were placed in groups of four and randomly allocated an endowment. Some players therefore started out richer than others in this game. Players were given the option to reduce (or increase) the income of others by purchasing negative (income-reducing) or positive (income-increasing) tokens and allocating these to other group members. In this setting, people allocated more negative tokens to the richest players and allocated more positive tokens to the poorest members of the group — suggesting that these behaviors were aimed at reducing outcome inequality. However, in this experiment, all four group members were able to purchase and allocate these tokens. Thus, it was impossible for players to predict how many tokens they would need to buy in order to achieve equal outcomes. Consequently, it is not possible to determine whether players adjusted investment in punitive behavior in order to achieve specific outcomes. Moreover, since initial payoff inequalities were exogenously determined rather than arising through some players defecting, the study could not test to what extent investment in income-reducing tokens was related to the target's behavior, as opposed to the outcome itself. In other words, since cooperation and defection were not possible in this game, any revenge-based motives of punishment could not be measured.

A more recent study by Houser and Xiao (2010) showed that players who were treated unfairly most commonly chose to punish as severely as possible and thus create inequality in their own favor. Although this seems to be more suggestive of punishment as a form of revenge rather than a fitness-leveler, it is important to take into account that in this study the severity of punishment chosen was not constrained by cost. In reality, imposing a larger cost on another individual is likely to also impose a larger cost on the punisher (Raihani & McAuliffe, 2012a). Since punishers have been shown to adjust their investment according to the costs associated with punishment (Anderson & Putterman, 2006; Bone, Silva, & Raihani, 2014; Carpenter, 2007; Nikiforakis & Normann, 2008; Ostrom, Walker, & Gardner, 1992), this creates a potentially important trade-off between maximizing income and achieving the desired punishment outcome.

The fitness-leveling hypothesis predicts that individuals should only invest in punishment that is more costly to the target than to the punisher, and is therefore able to reduce any existing disadvantageous inequality. Nevertheless, empirical work has demonstrated that individuals are prepared to invest in punishment that is equally costly to the punisher and the target (Anderson & Putterman, 2006; Carpenter, 2007; Egas & Riedl, 2008; Falk et al., 2005; Nikiforakis & Normann, 2008) — or even more costly to the punisher (Anderson & Putterman, 2006; Carpenter, 2007; Egas & Riedl, 2008) — and so is unable to re-establish equality. These findings suggest that punishers are not solely motivated by a desire to remove fitness differentials and support the idea that punishers might instead be motivated by a desire for revenge against defecting partners. The predictions of the two hypotheses also differ with respect to whether the defection was performed intentionally or not. Specifically, the revenge hypothesis predicts that punishment should be focused on those who impose harm intentionally and can therefore learn to avoid repeating the harmful behavior in the future. Conversely, punishment aimed at removing fitness differentials should be less sensitive (or insensitive) to intentionality since the primary function is to reduce inequality rather than change the target's behavior. Evidence from empirical studies provides some support for both hypotheses. Whilst several studies have shown that individuals will punish in response to unequal outcomes created at random or unintentionally (Cushman, Dreber, Wang, & Costa, 2009; Dawes et al., 2007; Falk, Fehr, & Fischbacher, 2008; Houser & Xiao, 2010; Kagel, Kim, & Moser, 1996; Yu, Calder, & Mobbs, 2014), individuals are significantly more likely to punish when unequal outcomes are created intentionally by the target (Falk et al., 2008; Houser & Xiao, 2010; Kagel et al., 1996).

Based on past research it is therefore unclear whether punishment is motivated by a desire for revenge or by a desire to equalize payoffs. We aimed to answer this question by investigating whether victims of cheats adjusted their investment in punishment in order to restore equality using a modified version of the game used by Raihani and McAuliffe (2012b). In the current study, one player could choose to steal $0.20 from their partner. Depending on the treatment, players interacting with a stealing partner experienced advantageous inequality, equal outcomes or varying levels of disadvantageous inequality. Players could punish stealing partners, but while some players had access to effective punishment (1:3 fee to fine) others could only use ineffective punishment (1:1 fee to fine). Players who had access to effective punishment could achieve equal outcomes by tailoring their investment in punishment: more extreme outcome inequality could be alleviated by investing more into punishment. However, under the ineffective punishment condition, increasing investment in punishment did not reduce inequality.

Although we suggest that revenge may serve a deterrent function, in the anonymous one-shot setting of our game, there is no scope for punishment to change the behavior of stealing partners (or bystanders). However, previous work has suggested that behavior may be constrained by psychological mechanisms that evolved in the context of non-anonymous repeated interactions and that responses that are attuned to these conditions may be invoked even in anonymous, one-shot settings (Ben-Ner & Putterman, 2000; Burnham & Johnson, 2005; Cosmides & Tooby, 1989; Delton, Krasnow, Cosmides, & Tooby, 2011; Hagen & Hammerstein, 2006; Hoffman, McCabe, & Smith, 1998; Johnson, Stopka, & Knights, 2003; Tooby, Cosmides, & Price, 2006). Thus, in our game a desire for revenge might reflect the desires of an evolved psychology that functions to deter cheats, even though this function is (due to the nature of the game) impossible to achieve. Nevertheless, we note that since deterrence is not the only possible function for this behavior we use the word 'revenge' in a purely descriptive sense.

The revenge hypothesis predicts that punishment will be used in both the ineffective and the effective punishment condition. Alternatively, if punishment is motivated by the desire to equalize outcomes,