Evolutionary Psychology

www.epjournal.net – 2013. 11(1): 186-200

Original Article

Who Punishes? Personality Traits Predict Individual Variation in Punitive Sentiment

S. Craig Roberts, Division of Psychology, School of Natural Sciences, University of Stirling, Stirling, UK. Email: craig.roberts@stir.ac.uk (Corresponding author).

Antonios Vakirtzis, Institute of Integrative Biology, University of Liverpool, Liverpool, UK.

Lilja Kristjánsdóttir, Institute of Integrative Biology, University of Liverpool, Liverpool, UK.

Jan Havlíček, Department of Zoology, Faculty of Science, Charles University, Prague, Czech Republic.

Abstract: Cross-culturally, participants in public goods games reward participants and punish defectors to a degree beyond that warranted by rational, profit-maximizing considerations. Costly punishment, where individuals impose costs on defectors at a cost to themselves, is thought to promote the maintenance of cooperation. However, despite substantial variation in the extent to which people punish, little is known about why some individuals, and not others, choose to pay these costs. Here, we test whether personality traits might contribute to variation in helping and punishment behavior. We first replicate a previous study using public goods scenarios to investigate effects of sex, relatedness and likelihood of future interaction on willingness to help a group member or to punish a transgressor. As in the previous study, we find that individuals are more willing to help related than unrelated needy others and that women are more likely to express desire to help than men. Desire to help was higher if the probability of future interaction is high, at least among women. In contrast, among these variables, only participant sex predicted some measures of punitive sentiment. Extending the replication, we found that punitive sentiment, but not willingness to help, was predicted by personality traits. Most notably, participants scoring lower on Agreeableness expressed more anger towards and greater desire to punish a transgressor, and were more willing to engage in costly punishment, at least in our scenario. Our results suggest that some personality traits may contribute to underpinning individual variation in social enforcement of cooperation.

Keywords: reciprocity, altruistic punishment, prisoner's dilemma, ultimatum game, evolutionary economics

Introduction

The evolution of cooperation is a problem of great interest not only to a wide range of evolutionary and behavioral disciplines, but also to policy makers and society at large (Hardin, 1968). Established partial solutions to this problem include kin altruism (Hamilton, 1964), reciprocal altruism (Axelrod and Hamilton, 1981; Trivers, 1971), reputation building (Alexander, 1987) and costly signaling (Zahavi, 1995). However, there remain several puzzling aspects in human cooperative behavior that seem to defy explanation by any of these mechanisms (e.g., Bowles and Gintis, 2002; Fehr, Firschbacher and Gächter, 2002; Fehr and Gächter, 2000, 2002). These behaviors arise in the context of experimental games which are intentionally structured so as to ensure that participants' behavior cannot be logically attributed to any of the standard explanations of cooperation. Players are total strangers to each other (excludes kin altruism), they make their decisions anonymously (excludes reputation building and costly signaling) and they only interact with each other once (excludes reciprocal altruism) although the latter condition may not perfectly capture real-world conditions, especially those in the ancestral past, where probability of future interaction would very rarely have been zero (see Burnham and Johnson, 2005; Hagen and Hammerstein, 2006; Trivers, 2004). Despite this, they still choose to reward participants and punish defectors in ways that are irrational from the viewpoint of calculating, self-interested agents who organize their behavior around income maximization (Fehr et al., 2002; Price, Cosmides, and Tooby, 2002). These puzzling results have been replicated with participants drawn from a wide range of cultures around the world (Henrich et al., 2006).

For example, the public goods paradigm (Egas and Riedl, 2008; Fehr and Gächter, 2000, 2002) typically involves participants who are given a number of monetary units (MU) at the start of the experiment, of which they can contribute any portion they want to a collective group project. Every MU invested by a player yields a payoff of less than 1 MU for every group member, while every MU withheld by a player yields a payoff of exactly 1 MU for that particular player and 0 MU for the other players. This payoff structure ensures that while all players would come out with a positive return to their investment if they invested their entire endowment to the common project, any single player can do even better by withholding funds from the project while reaping the benefits of the other players' contributions. All contributions are made simultaneously, they are anonymous, and the groups are broken up at the end of every round, such that no two participants ever play with each other more than once. At the end of every round participants are informed about the other members' contributions to the project (though not their identities) and can choose, at their own personal expense, to punish other members for not contributing to the project. For example participants might be allowed to give up 1 MU in order to deprive another player of 3 MU (Fehr and Gächter, 2002). Given the rules of the game a rational, self-interested player should never punish, yet players do routinely punish fellow players who make below-average contributions. Since all interactions are one-off and the punisher can never directly benefit from any subsequent change in the punished player's cooperative behavior, this type of behavior is often described as 'altruistic punishment' (Bowles and Gintis, 2002; Egas and Riedl, 2008; Fehr and Gächter, 2002). This sort of punishment has beneficial effects in sustaining high investment levels, as punished members raise their contributions in subsequent rounds, even though they interact with new players in each round. This surprising outcome stands in stark contrast to the no-punishment condition of the game, where players are not allowed to punish free-riders. In this condition cooperation gradually dwindles, with most participants eventually contributing nothing to the project (Fehr and Firschbacher, 2004; Fehr and Gächter, 2002).

Results like this have led to the emergence of 'strong reciprocity' as a general term to describe humans' tendency to punish defectors and reward co-operators beyond what is warranted by rational choice theory (Fehr et al., 2002; Gintis, 2000; Gintis, Henrich, Bowles, Boyd, and Fehr, 2008) and where other standard evolutionary explanations do not apply. There is yet no consensus as to the evolutionary significance of this phenomenon (Gintis et al., 2008; Kiyonari and Barclay, 2008; Price, 2008; Price et al., 2002; Sigmund, 2007), but one suggestion is that it may have evolved by enforcing cooperation between group members, with selection acting at the level of human groups (Boyd, Gintis, Bowles, and Richerson, 2003; Fehr et al., 2002; Gintis et al., 2008).

Previous studies indicate remarkable variation in individual engagement in both costly punishment and helping (Barclay, 2006; O'Gorman, Wilson, and Miller, 2005), and there is growing interest in the sources of this individual variation (Bergmüller, Schürch, and Hamilton, 2010). Indeed, this is true of behavior in experimental games in general. For example, there is widespread heterogeneity in rejection of offers in the ultimatum game (Fehr and Schmidt, 1999) and this appears to be under significant genetic influence (Wallace, Cesarini, Lichtenstein, and Johannesson, 2007). Wallace et al. (2007) argue that this variation may contribute to the mixed success that has met attempts to find satisfactory correlates of behavior in experimental games, and that further efforts should explore links between behavior in economic games and the significant source of variation in behavior that derives from personality traits. Likewise, Bergmüller et al. (2010) highlight the potential role of personality (or 'behavioral syndromes') in determining individual differences in cooperative behavior in animals.

The aim of this study was to explicitly investigate these possible links between personality traits and variation in helping and punishment behaviors. We framed our study within evolutionarily informed theories of personality that suggest that variance in human personality traits can be attributed to selective advantages of different responses to recurrent social problems faced by our ancestors (Buss, 1984, 1991; Michalski and Shackelford, 2010). Consequently, each trait can be considered to be a result of various trade-offs; for instance, individuals scoring high in extraversion report more sexual partners but also more injuries (Nettle, 2005). In this light, it is reasonable to question how personality traits help to shape individual behavior in the face of social dilemmas (Michalski and Shackelford, 2010). Here, motivated by growing recognition of the possibility that personality traits underpin at least some variation in cooperation in humans and animals (Wallace et al., 2007; Bergmüller et al., 2010), we investigate the role of personality in shaping behavior in social scenarios calling for cooperative responses or responses to a social transgression in a cooperative setting.

To do this, we set out to replicate and extend the design of O'Gorman et al. (2005) that tested individuals' desire to help or punish in various hypothetical scenarios. These scenarios elicited marked individual variation in willingness to help and punish. In their study, desire to help an unfortunate other was influenced by both relatedness and probability of future interaction, as predicted by evolutionary considerations. Individuals were more willing to help relatives than strangers, and more willing to help if there was high probability of future interactions. In contrast, desire to punish free-riders was moderated by neither of these variables, leading O'Gorman et al. to conclude that altruistic helping and punishing behavior were underpinned by different proximate psychological factors. Willingness to engage in punishment might therefore be influenced either by contextual variables not captured in O'Gorman et al.'s study, or by other causes of individual variation such as personality traits. Here, we specifically focused on this second potential source of variation – personality traits. Following general predictions about possible costs and benefits of each of the Big Five Model domains proposed by Nettle (2006), we considered how these domains might relate to tendency to engage in punishment. In particular, we predicted that willingness to punish might be specifically driven by variation along the Extraversion, Agreeableness, Conscientiousness and Neuroticism domains. It is well established that variation in these domains affects social interactions across various contexts. For example, Extraversion is intrinsically linked to engagement in social interactions. With regard to Agreeableness, the lower end of the dimension is characterized by descriptors such as hostility and spitefulness. Those scoring low on this dimension are less likely to compromise during inter-personal conflicts, while those scoring high receive higher levels of peer acceptance, have more mutual friends and better outcomes in cooperative tasks (Jensen-Campbell, Adams, Perry, Workman, Furdella, and Egan, 2002; Jensen-Campbell and Graziano, 2001; Koole, van den Berg, Vlek, and Hofstee, 2001). Nettle (2006) highlights that larger and looser social groupings will select for lower Agreeableness than smaller, closer ones. People scoring high on Conscientiousness tend to favor social norms (McCrae and John, 1992), exhibit strong moral principle (Nettle, 2006) and have higher peer acceptance (Jensen-Campbell and Malcolm, 2007). Finally, individuals who score highly on Neuroticism may experience anger more readily, which may lead to increased engagement in costly punishment (Seip, van Dijk, and Rotteveel, 2009).

Materials and Methods

Participants

We recruited 211 participants (104 male, 97 female, 10 not reported) who were visitors to a local science museum. Participants were aged between 16 and 78 years (*M* ± *SD*; 33.3 ± 12.6). This is a wider range than that in O'Gorman et al.'s (2005) study, allowing us to generalize beyond a narrow undergraduate sample.

Procedure

Each participant answered questions relating to two separate fictional scenarios (a helping and a punishment scenario) adapted from O'Gorman et al. (2005) (for a full description, see Appendix). The scenarios were presented in randomized order. Briefly, in the helping scenario, participants were asked to imagine that, together with nine others (either cousins or strangers), they had pooled their money to invest in the stock market, but the investment only broke even and one member had anticipated a profit to pay for emergency medical bills. This unfortunate person was either moving to another town or living in the same town, depending on the condition. This scenario therefore shared the punishment scenario's 2 (Relatedness) x 2 (Probability of future interaction) factorial design, with four dependent variables: sympathy for the person, desire to help the person, level of anger towards other group members that were unwilling to help the person (each on a 1 to 9 scale), and the maximum amount the participant would be willing to give to help the person.

The punishment scenario also involved the participant pooling investment funds with nine other people (cousins or strangers), but in this scenario one participant had cheated by contributing considerably less than the agreed-upon amount. While everyone had contributed £1000, the transgressor had contributed only £200, and, since the profits had been split equally, the transgressor ended up receiving £2160 more than he deserved. The transgressor was either moving to another town, or living in the same town, depending on the condition. This again yielded a 2 x 2 factorial design with four conditions in total. The dependent variables were responses to four items. Two of them recorded participants' anger towards, and desire to punish, the transgressor (on a nine point scale). There were also two items that asked participants to indicate the amount of money they thought the transgressor should pay and the amount they would personally be willing to spend in order to punish him/her.

Participants also provided their age and sex, and completed a brief measure of the Big Five personality factors, the Ten-Item Personality Inventory (TIPI), which includes two items per factor and is thus quick to administer (which suited the setting of this study) but nonetheless correlates well with both the longer Big-Five Inventory and the NEO Personality Inventory (Gosling, Rentflow, and Swann, 2003). In our data, internal consistency between items was as follows: extraversion (Cronbach's α = .67), agreeableness (α = .11), conscientiousness (α = .37), neuroticism (emotional stability in Gosling et al.; α = .45), openness (α = .26).

In analysis of the data, we again followed the approach of O'Gorman et al. (2005), using factorial ANOVA with relatedness and probability of future interaction as between-subject factors. We included participants' scores on the personality dimensions as covariates. Although the dependent variables were not always normally distributed, ANOVA is robust to this violation (Field, 2005). Data were analyzed in SPSS, version 18. Where appropriate, we include estimates of effect size, usually Cohen's *d,* for which *d* of .20, .50 and .80 are indicative of a small, medium and large effect, respectively (Cohen, 1992).

Results

Validity checking

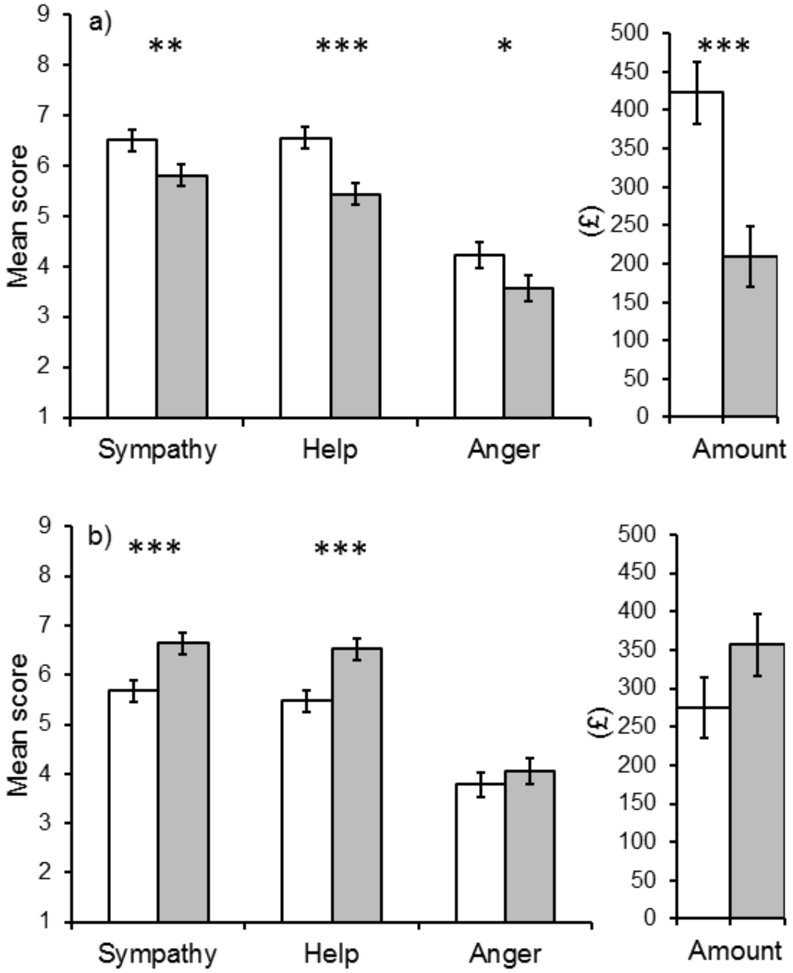
We first carried out a validity check of our scenario-setting, by analyzing the data as a simple replication of O'Gorman et al.'s study, without accounting for personality traits. Analysis of the helping scenario revealed a number of significant effects (in all cases, the overall ANOVA models were significant at α = .05). There were main effects of relatedness (see Figure 1a) with regard to sympathy for the person [$F(1,192) = 5.36$, p = .022, d = .33], desire to help [$F(1,190) = 12.77$, $p < .001$, d = .52] and amount of money participants would be willing to give [$F(1,174) = 14.39$, $p < .001$, d = .57], but not for anger at non-helpers [$F(1,185) = 3.21$, p = .08, d = .26]. There were also main effects of participant sex (see Figure 1b), with women expressing greater sympathy for the person [$F(1,192) = 9.97$, p = .002, d = .45] and desire to help [$F(1,190) = 11.37$, p = .001, d = .49], but there was no effect of sex on anger at non-helpers [$F(1,185) = .52$, p = .47, d = .11] or on the amount participants would be willing to give [$F(1,174) = 2.10$, p = .15, d = .22]. Contrary to O'Gorman et al. (2003), there was no main effect of probability of future interactions on helping, but there were several significant interaction effects with participant sex: analysis showed that while men's responses were relatively insensitive to the likelihood of future interaction, women tended to express more sympathy when future interactions were likely [$F(1,192) = 3.48$, p = .064, d = .27], were more willing to help [$F(1,190) = 4.14$, p = .043, d = .29], were more angry at non-helpers [$F(1,185) = 6.65$, p = .011, d = .38] and were willing to give more money to the unfortunate group member [$F(1,174) = 4.02$, p = .046, d = .31].

In contrast, in the punishment scenario, there were no significant main effects of relatedness, future interaction or sex across any of the four dependent variables, nor any significant interaction effects. The only factor that approached significance was a main effect of sex [$F(1,192) = 3.41$, p = .066, d = .26], but only in participants' desire to punish the cheater ($M \pm SE$; males: 5.4 ± 2.5, females: 4.7 ± 2.6). We further investigated the amount that participants would be willing to pay to punish because this variable was highly skewed, with 60% of participants indicating they would pay nothing at all. On this basis, we further tested this variable using binary logistic regression, comparing those who would pay nothing and those who were willing to pay something. This analysis revealed a main sex effect (*Wald's* χ^2 = 5.48, df = 1, p = .019) with men being more willing to pay to punish than women, and an interaction between sex and likelihood of future interaction (*Wald's* χ^2 = 3.99, df = 1, p = .046), such that women were especially reticent to pay when meeting again was more likely.

Effect of personality

To investigate the contributions of personality traits on responses to the scenarios, we re-ran the analyses as before, including participant sex, relatedness and likelihood of future interaction as factors, now also including as covariates the participants' scores on the Big Five personality dimensions.

Figure 1. Effects of (a) relatedness and (b) participant sex in the helping scenario. Data show mean (± *SE*) sympathy to an unfortunate group member, desire to help, anger at non-helpers, and the amount participants would be willing to give to help.



Note: In a), individuals are more sympathetic and willing to help cousins (open bars) than strangers (shaded bars), more angry at non-helpers if the unfortunate person is a cousin, and willing to pay more to help cousins than strangers. In b), male participants (open bars) are significantly less sympathetic and willing to help than female participants (shaded bars). **p* < .1, ** *p* <.05, ****p* < .01

In the helping scenario, the only dependent variable that was significantly predicted from any of the Big Five dimensions was sympathy for the unfortunate person. High scores on Agreeableness predicted increased sympathy towards the unfortunate group member (β = .31 ± .16, p = .047), with no other significant effects of personality dimension.

In the punishment scenario, in contrast, analyses revealed several significant relationships between participants' responses and scores on personality dimensions (see Table 1). In particular, participants who scored lower on the Agreeableness dimension expressed more anger towards (p = .008) and greater desire to punish (p < .0001) the transgressor, and thought the transgressor should pay more in reparation (*p* = .037). In addition, participants who scored higher on Conscientiousness expressed greater anger towards the transgressor (*p* = .033) and those scoring lower on Extraversion expressed greater desire to punish (*p* = .016). Finally, as before, we further tested effects of personality traits on those who were willing to pay at least something to punish the transgressor (a binarized version of the final dependent variable), using logistic regression. In this analysis, lower scores on the Agreeableness dimension also predicted willingness to pay to punish the transgressor (*p* = .045).

Table 1. Relationships between personality attributes and behavior directed towards a transgressor

	Anger		Desire		Amount he should pay back		Amount willing to pay to punish		Pay something to punish	
	B (*SE*)	*p*	*B* (*SE*)	*p*	*B* (*SE*)	*p*	*B* (*SE*)	*p*	*B* (*SE*)	*p*
E	-.083 (.096)	.324	**-.324** **(.133)**	**.016**	-41.5 (58.0)	.475	-22.2 (30.5)	.468	-.037 (.119)	.753
A	**-.324** **(.122)**	**.008**	**-.731** **(.169)**	**< .001**	**-157.4** **(74.8)**	**.037**	-9.06 (39.7)	.820	**-.307** **(.153)**	**.045**
C	**.227** **(.106)**	**.033**	.111 (.146)	.448	.94 (62.7)	.988	-.70 (33.8)	.984	-.156 (.130)	.233
N	-.047 (.103)	.645	-.099 (.142)	.486	3.96 (62.1)	.949	-27.1 (32.4)	.404	-.041 (.124)	.740
O	-.017 (.146)	.909	.153 (.202)	.448	51.9 (87.6)	.555	45.2 (49.9)	.984	-.080 (.178)	.654

Note: E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism; O = Openness; Significant effects are shown in bold

Discussion

Our results are broadly consistent with those of O'Gorman et al. (2005), who also found that participants were sensitive to context cues in the helping scenario but that these variables did not tend to predict behavior in the punishment scenario. Like O'Gorman et al., we found that relatedness cues predicted willingness to help, but not to punish. Effects of participant sex in our study were also in the same direction as those in O'Gorman et al.'s study: they found a near-significant tendency for women to express more sympathy for the group member in the helping scenario (this was significant in our study, as was desire to help), but in neither study did participant sex have effects on anger towards the transgressor or contributions to help. In the punishment scenario, O'Gorman et al. found that men expressed greater desire than women to punish the transgressor, while we found a near-significant difference in the same direction. In neither study were there significant sex differences in levels of anger, the amount that participants thought the transgressor should pay, nor in the amount they were willing to pay to punish (except where we binarized the data, which O'Gorman et al. did not do). As with O'Gorman et al., then, where sex differences did exist, men appeared more likely to engage in punishment than women. This could be because men tend to (and presumably did, in our evolutionary past) engage with larger groups on average than women, (e.g., Baumeister 2010), where the need for social enforcement of cooperation are greater.

The main difference between our results and those of O'Gorman et al. was that we did not find a main effect of probability of future interaction in the helping scenario, although we did find significant interactions with sex, such that this effect was more important in women than men. We do not know the reason for this difference, but it could be due to the combination of stronger effects in women and O'Gorman et al. having a higher proportion of women in their sample than we did. However, with this exception, our results replicated the main findings of the earlier study and confirm the validity of the scenario-setting in our study, thus enabling us to proceed to examine effects of personality traits.

As predicted, we found that engagement in punishment was correlated with personality, in particular with individuals' score on the Agreeableness dimension. Participants who scored low on Agreeableness expressed greater anger towards, and desire to punish, the transgressor. They also felt the transgressor should pay more in reparation, and they expressed greater willingness to pay at least something to punish the transgressor. Higher scores on Conscientiousness also predicted anger towards the transgressor, although it did not correlate with other measures. Furthermore, individuals scoring low on Extraversion also indicated a higher desire to punish the transgressor. Conscientiousness and Extraversion predicted single aspects of behavior only (respectively, anger towards and desire to punish the transgressor) and not measures of actual reparation or monetary will to punish. O'Gorman et al. also reported that while most of their participants expressed high levels of anger and willingness to punish a transgressor, only some were willing to actually engage in altruistic punishment. Thus, the effects of Conscientiousness and Extraversion extend to emotional reaction towards, but not action against, the transgressor, in contrast to individuals scoring low on Agreeableness.

The finding that people scoring lower on Agreeableness express greater anger towards defectors in a collaborative enterprise, and are more willing to punish them, is in line with their description as relatively suspicious, irritable and vengeful (e.g., McCraw and John, 1992; Costa and McCrae, 1992). Similarly, the finding that more conscientious people expressed greater anger towards defectors is not difficult to interpret. Goldberg (1993) describes them, among other things, as orderly, organized, systematic, punctual, economical and thrifty. Apart from the ethical dimension of the matter, defection by another group member would represent a considerable upset and disturbance to the conscientious person's organization and planning. We also found a negative relationship between extraversion and desire to punish. Introverts have been conceptualized, among other things, as timid and unassertive (Goldberg, 1992), traits that are difficult to reconcile with the results of this study, but a possible explanation might be found in their lower reliance on social interactions, and consequently increased tolerance of dissatisfaction or even hostility on the part of others. Introverts might therefore be more willing to assume the social costs of retaliation (Janssen and Bushman, 2008). This line of reasoning suggests that the costs of punishment are considerable, and can potentially disrupt extraverts' fabric of social relations.

It was perhaps surprising that individual variation in desire to punish, and anger at transgressors in the collaborative venture, were not predicted by participants' scores on the Neuroticism dimension. Individuals who score highly on this dimension tend to experience emotions such as anger more readily, and anger has been implicated in the tendency to engage in costly punishment (Seip, van Dijk, and Rotteveel, 2009), as well as competitive attack responses in a wartime prisoner's dilemma game (Kassinove, Roth, Owens, and Fuller, 2002). In one of Seip, van Dijk, and Rotteveel's (2009) studies, more punishment was elicited in participants who had been primed with feelings of anger (by recalling and describing an autobiographical episode involving anger) compared to a control group. It thus appears likely that the emotion of anger is important in driving punishment behavior, and in our study, level of anger at the transgressor was highly correlated with desire to punish (r = .545, p < .001). However, while anger as an emotion is not restricted to individuals scoring highly on neuroticism, one would expect such people to display higher than average levels of anger. Thus, the lack of relationship between punishment and Neuroticism indicates that the links between personality traits and punishment behavior extend beyond proneness to anger.

Before we leave this discussion, however, one should note that the tool we used to assess personality dimensions might in fact underestimate the effects under investigation. Due to time constraints on our participants (members of the public passing through a science museum), we employed a ten-item personality (TIPI) questionnaire, assessing individual scores on each of the five personality dimensions on the basis of responses to only two items. Although scores on the TIPI correlate well with scores generated by more detailed instruments (Gosling et al. 2003), its internal consistency is lower than typically found in such instruments (in both our sample and that originally reported by Gosling et al.). Thus, although the fact that we found significant relationships between personality and punishment variables when using this tool implies a robust effect, future studies should consider use of a more detailed personality questionnaire.

A potential criticism of our study is that it relies on responses to hypothetical scenarios and involves no actual monetary (or other) cost to participants. In light of the findings we describe here, it would certainly be interesting to examine whether and how the links between personality and punishment hold in an experimental setting. However, as O'Gorman et al. (2005) discussed at some length, the scenarios presented in their study and also used here were designed to resemble real-world interactions, produced comparable results to experiments on altruistic punishment, and strongly engaged the participants. Like them, we found marked variation between participants in levels of anger, desire to punish and willingness to punish at their own expense. Even though they were fictional scenarios and did not involve actual cost to participants, a very high proportion (60%) indicated they would not pay anything to punish, indicating they were treating the scenario seriously and behaving as one would expect in the real world. Furthermore, the overall pattern of our results map very closely onto O'Gorman et al.'s, indicating reliability of the scenarios, and both sets of results produce results that are directly predicted by evolutionary principles such as kin selection and game theory. For these reasons, we concur with O'Gorman et al. that the effects elicited by these scenarios are unlikely to be an artefact of the procedure; on the contrary, we believe it is more likely than not that they reflect how people will respond to actual social injustices encountered in the real world.

Although here we have concentrated solely on psychological traits, studies have also begun to address the possible utility of biological markers in predicting punishing behavior. It has been found, for example, that men with higher levels of testosterone are more likely to punish selfish offers in the ultimatum game (Burnham, 2007). Along with such measures, our results suggest that personality may play a critical role in determining individual variation in willingness to engage in punishment. This provides at least a partial explanation for apparently irrational behavior in experimental games: because personality types are relatively stable, patterns of behavior associated with certain personality dimensions could render individuals susceptible to engaging in enigmatic costly punishment, even in one-shot interactions.

Acknowledgments: We are grateful for permission to carry out this research at the World Museum, Liverpool, and thank all the participants for taking part. We also thank April Bleske-Rechek for constructive comments. JH is supported by Czech Science Foundation grant (GACR P407/10/1303) and Charles University Research Centre (UNCE 204004).

Received 19 August 2011; Revision submitted 4 May 2012; Revision submitted 15 November 2012; Accepted 16 November 2012

References

Alexander, R. D. (1987). *The biology of moral systems*. New York: Aldine de Gruyter.

Axelrod, R., and Hamilton, W. D. (1981). The evolution of cooperation. *Science, 211*, 1390-1396.

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior, 27*, 325-344.

Baumeister, R. F. (2010). *Is there anything good about men?* New York: Oxford University Press.

Bergmüller, R., Schürch, R., and Hamilton, I. (2010). Evolutionary causes and consequences of consistent individual variation in cooperative behaviour. *Philosophical Transactions of the Royal Society B, 365*, 2751-2764.

Bowles, S., and Gintis, H. (2002). Homo reciprocans. *Nature, 415*, 125-128.

Boyd, R., Gintis, H., Bowles, S., and Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences, USA, 100*, 3531-3535.

Burnham, T. C. (2007). High testosterone men reject low ultimatum game offers. *Proceedings of the Royal Society of London B, 274*, 2327-2330.

Burnham, T., and Johnson, D. D. P. (2005). The evolutionary and biological logic of human cooperation. *Analyse and Kritik*, *27*, 113-135.

Buss, D. M. (1984). Evolutionary biology and personality psychology: Toward a conception of human nature and individual differences. *American Psychologist, 39*, 1135-1147.

Buss, D. M. (1991). Evolutionary personality psychology. *Annual Review of Psychology, 42*, 459-491.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Costa, P. T., Jr, and McCrae, R. R. (1992). *NEO-PI-R: Professional manual*. Odessa, FL: Psychological Assessment Resources.

Egas, M., and Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society of London B, 275*, 871-878.

Fehr, E., and Firschbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences, 8*, 185-190.

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

Fehr, E., and Gächter, S. (2000). Cooperation and punishment in public goods experiments. *The American Economic Review, 90*, 980-994.

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

Fehr, E., and Schmidt, K. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics, 114*, 817-868.

Field, A. (2005). *Discovering statistics using SPSS*. London: Sage.

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology, 206*, 169-179.

Gintis, H., Henrich, J., Bowles, S., Boyd, R., and Fehr, E. (2008). Strong reciprocity and the roots of human morality. *Social Justice Research, 21*, 241-253.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4*, 26-42.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*, 26-34.

Gosling, S. D., Rentflow, P. J., and Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*, 504-528.

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

Hamilton, W. D. (1964). The genetical evolution of social behavior. *Journal of Theoretical Biology, 7*, 1-52.

Hardin, G. (1968). The tragedy of the commons. *Science, 162*, 1243-1248.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., . . . Ziker, J. (2006). Costly punishment across societies. *Science, 312*, 1767-1770.

Janssen, M. A., and Bushman, C. (2008). Evolution of cooperation and altruistic punishment when retaliation is possible. *Journal of Theoretical Biology, 254*, 541-545.

Jensen-Campbell, L. A., and Graziano, W. G. (2001). Agreeableness as a moderator of interpersonal conflict. *Journal of Personality, 69*, 323-362.

Jensen-Campbell, L. A., Adams, R., Perry, D. G., Workman, K. A., Furdella, J. Q., and Egan, S. K. (2002). Agreeableness, extraversion, and peer relations in early adolescence: Winning friends and deflecting aggression. *Journal of Research in Personality, 36*, 224-251.

Jensen-Campbell, L. A., and Malcolm, K. T. (2007). The importance of conscientiousness in adolescent interpersonal relationships. *Personality and Social Psychology Bulletin, 33*, 368-383.

Kassinove, H., Roth, D., Owens, S., and Fuller, J. (2002). Effects of trait anger and anger expression style on competitive attack responses in a wartime prisoner's dilemma game. *Aggressive Behavior, 28*, 117-125.

Kiyonari, T., and Barclay, P. (2008). Cooperation in social dilemmas: Free riding may be thwarted by second-order reward rather than by punishment. *Journal of Personality and Social Psychology, 95*, 826-842.

Koole, S. L., Jager, W., van den Berg, A. E., Vlek, C. A. J., and Hofstee, W. K. B. (2001). On the social nature of personality: Effects of extraversion, agreeableness, and feedback about collective resource use on cooperation in a resource dilemma. *Personality and Social Psychology Bulletin, 27*, 289-301.

McCrae, R. R., and John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60*, 175-215.

Michalski, R. L., and Shackelford, T. K. (2010). Evolutionary personality psychology: Reconciling human nature and individual differences. *Personality and Individual Differences, 48*, 509-516.

Nettle, D. (2005). An evolutionary perspective on the extraversion continuum. *Evolution and Human Behavior, 26*, 363-373.

Nettle, D. (2006). The evolution of personality variation in humans and other animals. *American Psychologist, 61*, 622-631.

O'Gorman, R., Wilson, D. S., and Miller, R. R. (2005). Altruistic punishing and helping differ in sensitivity to relatedness, friendship, and future interactions. *Evolution and Human Behavior, 26*, 375-387.

Price, M. E. (2008). The resurrection of group selection as a theory of human cooperation. *Social Justice Research, 21*, 228-240.

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

Seip, E., van Dijk, W., and Rotteveel, M. (2009). On Hotheads and Dirty Harries: The primacy of anger in altruistic punishment. *Values, Empathy, and Fairness Across Social Barriers, 1167*, 190-196.

Sigmund, K. (2007). Punish or perish? Retaliation and collaboration among humans. *Trends in Ecology and Evolution, 22*, 593-600.

Trivers, R. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology, 46*, 35-57.

Trivers, R. (2004). Mutual benefits at all levels of life. *Science, 304*, 964-965.

Wallace, B., Cesarini, D., Lichtenstein, P., and Johannesson, M. (2007). Heritability of ultimatum game responder behavior. *Proceedings of the National Academy of*

Sciences USA, 104, 15631-15634.

Zahavi, A. (1995). Altruism as a handicap: The limitations of kin selection and reciprocity. *Journal of Avian Biology, 26*, 1-3.

Appendix

Scenarios presented to participants

Two scenarios were presented to each participant in randomised order. Sections in *italics* and [square brackets] denote alternative parts of the scenario presented to different participants, altered to manipulate genetic relatedness and probability of future interaction. The titles to the two scenarios (helping scenario and punishment scenario), presented here for the reader's benefit, were not included in the presented scenario.

A. Helping scenario.

Suppose that you and *your nine cousins* [nine strangers] in your town decide to pool your money to invest in the stock market. Pooling the money allows the group to qualify for lower transaction costs than if each person invests separately. Each person aggress to contribute £1000 to the pool and to equally share the profits or losses. The stocks only break even after only one year. One investing member had anticipated a profit to pay for emergency medical bills.

All of you live in the same town and this person is likely to [This person is moving to another town and is unlikely to] associate with you and *your cousins* [or your friends] in the future. How much are you willing to contribute towards helping?

1) How sorry would you feel towards this person? Please answer on a scale from 1 (not at all sorry) to 9 (extremely sorry).

2) How much would you like to help this person? Please answer on a scale from 1 (no interest in helping) to 9 (extremely interested in helping).

3) What is the most you would be willing to give to help this person?

4) Some of the other group members want to help the person, but others seem unwilling to help if it costs them anything. How angry would you feel towards those who do not want to help? Please answer on a scale from 1 (not at all angry) to 9 (extremely angry).

B. Punishment scenario.

Suppose that you and *your nine cousins* [nine strangers] in your town decide to pool your money to invest in the stock market. Pooling the money allows the group to qualify for lower transaction costs than if each person invests separately. Each person agrees to contribute £1000 to the pool and to equally share the profits or losses. The stocks do very well and triple in value after only one year. Just before you meet to divide the profits, you discover that the person who volunteered to keep the books only invested £200, changing the records so that others would not notice. You do some calculations and determine the following facts:

1. The total amount invested was £9,200 which tripled in value to £27,600.
2. This was divided equally among all 10 friends to yield £2,760 for each person.
3. The person who contributed £200 should have received only £600 and therefore received £2160 more than deserved.
4. Everyone else should have received £3000, or £240 more than they actually got.

All of you live in the same town and this person is likely to [The person is moving to another town and is unlikely to] associate with you *or the other members of the club* [and your cousins] in the future. How are you going to act?

1) How angry would you feel toward this person? Please answer on a scale from 1 (not at all angry) to 9 (extremely angry).

2) How much would you like to punish this person? Please answer on a scale from 1 (no interest in punishing to 9 (extremely interested in punishing).

3) Although punishment can take many forms, if you think of it as an amount in pounds, how much do you think this person should pay for what he did?

4) Punishing this person can take many forms, but if you think of it as an amount in pounds, what is the most you would be willing to pay to punish this person?