Evolution and Human Behavior 37 (2016) 255–262

Contents lists available at ScienceDirect

Evolution and Human Behavior

journal homepage: www.ehbonline.org

Original Article

Generosity is a sign of trustworthiness—the punishment of selfishness is not

Wojtek Przepiorka [a,b,*], Ulf Liebe [c]

[a] Department of Sociology, Utrecht University, Padualaan 14, Utrecht 3584 CH, The Netherlands
[b] Nuffield College, New Road, Oxford OX1 1NF, United Kingdom
[c] Institute of Sociology, University of Bern, Fabrikstrasse 8, CH-3012 Bern, Switzerland

ARTICLE INFO

Article history:
Initial receipt 27 July 2015
Final revision received 28 December 2015

Keywords:
Peer-punishment
Third-party punishment
Signaling theory
Generosity
Trust
Trustworthiness

ABSTRACT

Peer-punishment is an important determinant of cooperation in human groups. It has been suggested that, at the proximate level of analysis, punitive preferences can explain why humans incur costs to punish their deviant peers. How punitive preferences could have evolved in humans is still not entirely understood. A possible explanation at the ultimate level of analysis comes from signaling theory. It has been argued that the punishment of defectors can be a type-separating signal of the punisher's cooperative intent. As a result, punishers are selected more often as interaction partners in social exchange and are partly compensated for the costs they incur when punishing defectors. A similar argument has been made with regard to acts of generosity. In a laboratory experiment, we investigate whether the punishment of a selfish division of money in a dictator game is a sign of trustworthiness and whether punishers are more trustworthy interaction partners in a trust game than non-punishers. We distinguish between second-party and third-party punishment and compare punitive acts with acts of generosity as signs of trustworthiness. We find that punishers are not more trustworthy than non-punishers and that punishers are not trusted more than non-punishers, both in the second-party and in the third-party punishment condition. To the contrary, second-party punishers are trusted less than their non-punishing counterparts. However, participants who choose a generous division of money are more trustworthy and are trusted more than participants who choose a selfish division or participants about whom no information is available. Our results suggest that, unlike for punitive acts, the signaling benefits of generosity are to be gained in social exchange.

1. Introduction

In the last three decades, the literature on peer-punishment as a mechanism to sustain cooperation in humans has thrived (Oliver, 1980; Axelrod, 1986; Boyd & Richerson, 1992). In situations in which individuals have an incentive to free ride on others' cooperative efforts, the presence of group members who punish free riders at an immediate cost to themselves can promote and maintain cooperation in the group (Ostrom, Walker, & Gardner, 1992; Fehr & Gächter, 2002; Gürerk et al., 2006). If the benefits of a cooperative environment outweigh the costs of maintaining a credible punishment threat, then peer-punishment is both rational and fitness enhancing and thus can be explained both at the proximate and ultimate level of analysis, respectively (Gächter et al., 2008; Raihani & Bshary, 2011; Przepiorka & Diekmann, 2013; Roberts, 2013). However, costly peer-sanctioning has been observed in one-time-only encounters between unrelated individuals both in the lab (Fehr & Fischbacher, 2004; Diekmann & Przepiorka, 2015) and in the field (Henrich et al., 2006; Balafoutas, Nikiforakis, & Rockenbach, 2014; Diekmann, Jann, Przepiorka, & Wehrli, 2014). In these situations, the benefits of peer-punishment are unlikely to outweigh the costs and, therefore, peer-punishment cannot be readily explained from within the rational choice and the individual-selectionist framework, respectively (Hamilton, 1963; Trivers, 1971; Becker, 1976). Punitive preferences have been proposed as a proximate explanation for why humans sanction their peers even in situations in which they incur a net loss (Gintis, 2000; Fehr, Fischbacher, & Gächter, 2002), and there is an ongoing debate about the function such punitive preferences evolved to fulfill (Sigmund, 2007; Dreber, Rand, Fudenberg, & Nowak, 2008; Baumard, 2010; Boyd, Gintis, & Bowles, 2010; Raihani & Bshary, 2011; West, El Mouden, & Gardner, 2011; Barclay, 2012; Guala, 2012; Krasnow, Cosmides, Pedersen, & Tooby, 2012).

One explanation for the evolution of punitive preferences which has received little attention comes from signaling theory (Spence, 1974; Zahavi, 1975, 1977; Bliege Bird & Smith, 2005; Gambetta, 2009). It has been argued that pro-social acts can function as a type-separating signal of an individual's unobservable quality, if this quality is causally related to the individual's ability to cooperate (Zahavi, 1995; Gintis, Smith, & Bowles, 2001). This argument has received empirical support. It has

* Corresponding author. Department of Sociology, Utrecht University, Padualaan 14, Utrecht 3584 CH, The Netherlands. Tel.: +31 30 253 3467.
E-mail address: w.przepiorka@uu.nl (W. Przepiorka).

http://dx.doi.org/10.1016/j.evolhumbehav.2015.12.003
1090-5138/

been shown that generosity can be a type-separating signal of an individual's trustworthiness (Barclay, 2004; Fehrler & Przepiorka, 2013; Gambetta & Przepiorka, 2014; Bliege Bird & Power, 2015) and that the signaling benefits of altruistic acts which accrue in social exchange can ease the conditions under which other-regarding preferences can evolve (Delton, Krasnow, Cosmides, & Tooby, 2011; Fehrler & Przepiorka, 2013). Since peer-punishment is often conceived as prosocial or even altruistic (Fehr & Gächter, 2002; Boyd, Gintis, Bowles, & Richerson, 2003; de Quervain et al., 2004; Fowler, 2005), it has been argued that peer-punishment too could work as a type-separating signal of the punisher's cooperative intent (Gintis et al., 2001). There is further theoretical support for this argument.

Many evolutionary models which show that punishment can promote the evolution of cooperation devise conditions under which cooperators who also punish defectors constitute an evolutionary stable strategy (Boyd et al., 2003; Hauert, Traulsen, Brandt, Nowak, & Sigmund, 2007; Helbing, Szolnoki, Perc, & Szabó, 2010; dos Santos, Rankin, & Wedekind, 2011). Under these conditions, punishment and (first-order) cooperation will be correlated and thus the former will be a reliable sign of the latter. Although there is empirical evidence in support of a signaling account of peer-punishment (Barclay, 2006; Kurzban, DeScioli, & O'Brien, 2007; Nelissen, 2008; Simpson, Harrell, & Willer, 2013), there is also evidence opposing or not supporting it (Rockenbach & Milinski, 2011; Pedersen, Kurzban, & McCullough, 2013; Balafoutas et al., 2014; FeldmanHall, Sokol-Hessner, Bavel, & Phelps, 2014; Gordon, Madden, & Lea, 2014). Based on a comprehensive review of this literature, a more elaborate argument recently emerged which tries to pin down the conditions under which we can expect peer-punishment to be a sign of a punisher's cooperative intent (Raihani & Bshary, 2015a).

The conceptual framework put forward by Raihani and Bshary (2015a) is informed by the growing literature investigating the proximate mechanisms behind individuals' punitive acts (Xiao & Houser, 2005; Jordan, McAuliffe, & Rand, 2015; Bone & Raihani, 2015). Peer-punishment can be triggered by different motives across different contexts (Leibbrandt & López-Pérez, 2012). Thus, punitive acts may be ambiguous in the information they convey about punishers' underlying motivations (Brañas-Garza, Espín, Exadaktylos, & Herrmann, 2014; Raihani & Bshary, 2015a). In particular, it has been argued theoretically and shown empirically that different motives might trigger peer-punishment in so-called second-party and in third-party punishment situations (Fehr & Fischbacher, 2004; Henrich et al., 2006; Carpenter & Matthews, 2009; Marlowe et al., 2011; Rockenbach & Milinski, 2011; Leibbrandt & López-Pérez, 2012; FeldmanHall et al., 2014; Gummerum & Chu, 2014; Harris, Herrmann, Kontoleon, & Newton, 2015). Punishing a deviant peer on one's own behalf (second-party punishment) is more likely to be motivated by vengefulness and thus more likely to be perceived as such by an observer (Marlowe et al., 2011; Rockenbach & Milinski, 2011). Punishing a deviant peer on the part of another "victim" (third-party punishment) is more likely to be motivated by the normative desire to establish justice and more likely to be perceived as such by an observer (Willer, Kuwabara, & Macy, 2009; Simpson et al., 2013; FeldmanHall et al., 2014).

1.1. Research question and hypotheses

Here we address the question whether peer-punishment can function as a type-separating signal of a punisher's cooperative intent. There are two necessary conditions for a signaling account of peer-punishment to be plausible. First, punitive preferences and cooperative intent must be positively related. Second, observers must infer cooperative intent from punitive acts. We conduct a laboratory experiment with economic games to test whether these two conditions are met. In our experiment, we measure subjects' punitive preferences in terms of their decisions to punish another subject for a selfish (i.e. self-regarding) division of money in a binary dictator game, we measure subjects' cooperative intent in terms of their trustworthiness as second movers in a trust game, and we measure whether subjects infer trustworthiness from punitive acts by these subjects' trust as first movers in the trust game. Trustworthiness is a concept widely used in the social sciences and stands for the cooperative intent of the second-moving party in social exchange (Coleman, 1990; Hardin, 2002; Gambetta & Hamill, 2005; Fehr, 2009). Our first two hypotheses can be stated as follows:

H1. *Actors who punish selfish behavior are more trustworthy than actors who do not punish selfish behavior.*

H2. *Actors who punish selfish behavior are trusted more than actors who do not punish selfish behavior.*

Based on the literature cited above, we expect that third-party punishment is a better sign of trustworthiness than second-party punishment, because we expect that a sense of justice sustains trustworthiness better than vengefulness does (Marlowe et al., 2011; Raihani & Bshary, 2015a). However, these two motives cannot be readily separated. For example, it cannot be ruled out *a priori* that a sense of justice will sometimes trump vengefulness in motivating second-party punishment. In the context of kinship relations and close friendships, vengefulness may also trump a sense of justice in motivating third-party punishment. We reduce the likelihood of vengefulness to motivate third-party punishment to a minimum by design. In our laboratory experiment, subjects interact with each other from behind their computer screens while sitting in isolated cubicles; the anonymous environment reduces the ability of third-party observers to empathize with recipients of a selfish division in the dictator game. Accordingly, our next two hypotheses are as follows:

H3. *The positive relation hypothesized under H1 is stronger with regard to third-party punishment than with regard to second-party punishment.*

H4. *The positive relation hypothesized under H2 is stronger with regard to third-party punishment than with regard to second-party punishment.*

Finally, we compare the information punitive acts convey with the well-established finding that generosity is positively related with trustworthiness and observers infer trustworthiness from acts of generosity. We call the more equal division of money in our binary dictator game "generous," although it need not be motivated by generosity alone, but could also be motivated by a sense of fairness or the adherence to a social norm for sharing; what matters is that all these motives too can sustain trustworthiness (Gambetta & Przepiorka, 2014). Hence, our last two hypotheses can be stated as follows:

H5. *Actors who are generous are more trustworthy than actors who are selfish.*

H6. *Actors who are generous are trusted more than actors who are selfish.*

To our knowledge, this is the first experimental study to directly compare punitive acts and acts of generosity as signs of trustworthiness. Given that both punitive acts and acts of generosity are important elements of human sociality, their relative importance as signs of trustworthiness will emerge from the direct comparison. At the time we conducted our experiment, we did not have any expectations as to whether generosity or punishment would prove to be the *better* sign of trustworthiness.

2. Materials and methods

2.1. Experimental games

We use the binary dictator game with second-party punishment (DG2P) and third-party punishment (DG3P) to measure subjects'

generosity and punitive preferences (Fehr & Fischbacher, 2004), and we use the trust game (TG) to measure their trust and trustworthiness (Dasgupta, 1988; Kreps, 1990). The three games are illustrated in Fig. 1a, b and c, respectively.

In the DG2P (Fig. 1a), two subjects are randomly paired, they are randomly assigned to be person A or B, they are endowed with CHF (i.e. Swiss Franc) 8 each, and person A can decide how to divide an additional amount of CHF 15 between him- or herself and person B. Person A can either keep CHF 8 and give CHF 7 to B (*generous* division), or keep CHF 12 and give only CHF 3 to B (*selfish* division). Person A decides knowing that next B can either *accept* the division or *punish* person A. If person B decides to punish A, person B incurs a punishment cost of CHF 3 and A is deducted CHF 9. The DG3P (Fig. 1b) is the same as the DG2P except for the fact that a third person C, who is endowed with CHF 14, observes person A's decision, and person A decides knowing that next C can either *accept* the division or *punish* person A. If person C decides to punish A, person C incurs a cost of CHF 3 and A is deducted CHF 9. Person B does not have the possibility to punish A, and person B's payoff is unaffected by C's decision. In the TG (Fig. 1c), two subjects are randomly paired and are randomly assigned to be person X or Y. Persons X and Y are endowed with CHF 8 each, and X can decide whether to equally *split* an additional amount of CHF 8 between him- or herself and person Y, or *send* the entire amount to Y. If person X decides to split, X and Y earn CHF 4 each. If person X decides to send, the amount is doubled and Y can decide whether to equally split the CHF 16 and *return* half to person X, or *keep* everything leaving nothing for X.

We employ the strategy method to elicit subjects' decisions (Selten, 1967; Brandts & Charness, 2011). That is, subjects make their decisions contingent on all relevant decisions their interaction partners *could* have made previously. For example, in the DG3P (see Figure 1b), a subject decides as person C both for the case that his or her interaction partner chose the generous division and for the case that his or her interaction partner chose the selfish division as person A (see the next section for details). However, the decision situations presented to subjects are not entirely hypothetical; they materialize based on subjects' and their interaction partners' actual decisions and determine subjects' earnings in the experiment. Thus, using the strategy method has the advantage that incentivized decisions can be obtained from subjects in all possible decision situations. At the same time, the strategy method might induce so-called experimenter demand effects; the different decision situations presented to subjects may clue subjects on the objective of the experiment and affect their behavior accordingly (Zizzo, 2010). Although we cannot entirely exclude that our using the strategy method induces experimenter demand effects, by using the strategy method, we give our hypotheses the best shot; if our hypotheses remain unsupported by the empirical evidence, they will hardly find support in experiments less susceptible to experimenter demand effects.

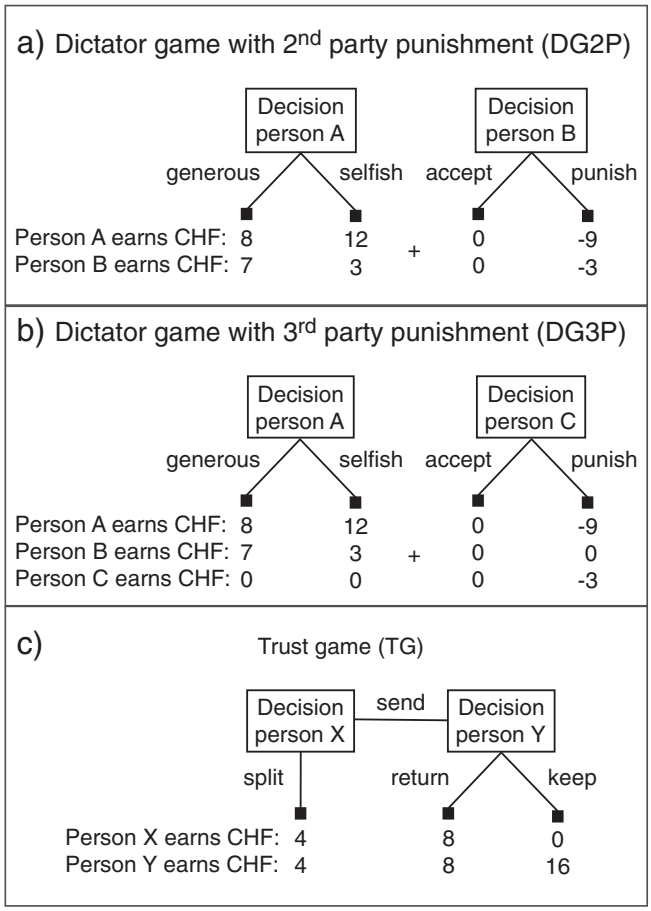

Fig. 1. The dictator game with 2nd party punishment (a) and 3rd party punishment (b) is used to measure subjects' generosity and punitive preferences, and the trust game (c) is used to measure subjects' trust and trustworthiness.

2.2. Experimental design

In the first part of the experiment (see Table 1), subjects decide as dictators (i.e. person A) and punishers (i.e. person B or person C) in a DG2P or DG3P; in the second part subjects also decide as trusters and trustees in a TG (i.e. person X and person Y). This design feature allows us to estimate the extent to which generosity, punishment and trustworthiness are related.

The second part of our experiment comprises three conditions (Table 1). In the control condition ("TG no info"), subjects decide as trusters in the TG without receiving any information about the trustees. In one treatment condition ("TG info DG2P"), subjects decide as trusters contingent on how generous the trustee was toward another person in the DG, or whether the trustee punished his interaction partner for a selfish division in the DG. In the second treatment condition ("TG info DG3P"), subjects decide as trusters contingent on how generous the trustee was toward another person in the DG, or whether the trustee (as a third party) punished another person for a selfish division in the DG. This design feature allows us to disentangle the trustworthy-making properties of generosity and peer-punishment on the one hand, and of second-party and third-party punishment on the other hand.

Moreover, our design rules out any possibilities for strategic reputation building (Engelmann & Fischbacher, 2009). First, subjects interact

Table 1
Experimental design.

	Experimental conditions, part I	
Stage	DG2P (n = 93)	DG3P (n = 93)
1	Decision as person B	Decision as person C
2	Decision as person A	Decision as person A
3	Belief elicitation	Belief elicitation

	Experimental conditions, part II		
Stage	TG info DG2P (n = 54)	TG no info (n = 77)	TG info DG3P (n = 55)
1	Decision as person Y	Decision as person Y	Decision as person Y
2	Decision as person X cond. on what person Y did as A or B in Part I	Decision as person X not knowing what person Y did in Part I	Decision as person X cond. on what person Y did as A or C in Part I
3	Belief elicitation	Belief elicitation	Belief elicitation

Notes: Part I: In the dictator game with 2nd party punishment (DG2P) and in the dictator game with 3rd party punishment (DG3P), subjects decide as dictators (i.e. person A) and as potential punishers (i.e. person B or C). Part II: In the trust game (TG), subjects also decide as trusters (i.e. person X) and trustees (i.e. person Y). The decision sequences in the two parts of our experiment were reversed. That is, subjects first decided as potential punishers and trustees before they decided as dictators and trusters, respectively. We reversed the decision sequences because we wanted to make it easier for subjects to put themselves in the shoes of potential punishers and trustees when deciding as dictators and trusters, respectively.

in the TG only once. Therefore, trusters in the two treatment conditions can infer trustees' trustworthiness (or untrustworthiness) based alone on what the trustees did in the first part of the experiment. Second, although subjects are told from the start that the experiment comprises two parts, they are only told at the end of the first part what the second part is about. With this design feature, we increase the likelihood that subjects make their decisions naturally, i.e. without anticipating the future information value of their choices. This also means that actions do not have to be costly to be type separating (Gambetta & Przepiorka, 2014). If subjects do not anticipate the information value of their actions, those who punish a selfish division of money, for instance, do so because they are genuinely inclined to do so, whereas those who lack such preferences do not punish because they do not expect to gain from it. Consequently, even if the punishment of a selfish division of money is not very costly in monetary terms, its occurrence (or absence) may well tell observers something about subjects' motives for acting in this way free from strategic considerations.

Finally, we also measure trusters' beliefs regarding trustees' trustworthiness in the TG. Trusters must believe that the trustees they transfer or refuse to transfer money to in the TG are trustworthy or untrustworthy, respectively. Only then we can be sure that trusters' behavior is not merely motivated by a desire to reward or punish the good or bad deeds, respectively, trustees did in the first part of the experiment (Raihani & Bshary, 2015b). Trust (or distrust) can be assumed to be motivated by pure self-interest, whereas rewards and punishments cannot and hence would remain in need of an ultimate explanation (Fehrler & Przepiorka, 2013).

2.3. Experimental procedure

We conducted six experimental sessions with 24–35 participants per session ($N = 186$ participants in total). Subjects were mostly undergraduate students from different departments at University of Zurich and ETH Zurich, 51% were female, and they were 23.6 years old on average (sd = 6.71). An experimental session lasted 75 minutes and subjects earned CHF 38 (≈USD 30) on average, including the show-up payment of CHF 10. Before the six experimental sessions, we also conducted one pilot session with 23 participants. Since we changed the payoffs at the punishment stage after the pilot session, we do not include the data from the pilot session in our analyses. The experiment was conducted in the Decision Science Laboratory of ETH Zurich (DeSciL), and we used the software zTree (Fischbacher, 2007) to program and run the experiment. In accordance with the lab's operational rules, no deception was used.

Upon arrival in the laboratory, participants were randomly assigned to experimental conditions DG2P or DG3P (Table 1), and they received condition specific part I instructions on paper. Subjects learned that the experiment comprised two parts, that they would receive the instructions for the second part only after the first part, that their decisions were anonymous, that their earnings would correspond to the sum they earned in both parts plus a show-up payment of CHF 10, and that they would receive their earnings at the end of the experiment from a person who was not involved in the implementation of the experiment. Most importantly, the instructions explained the decision situation step by step and how subjects' earnings depend on the decisions they would make and the decisions other participants would make. After reading the instructions, subjects were asked six control questions about the instructions on the screen. After answering them, all control questions were read out loud and the correct answers were explained to all subjects before the first part of the experiment started.

In part I, all subjects in the DG2P condition first decided as a person B and thereafter they decided as a person A. As person B, they decided whether to accept the division or punish A for both the case that A had chosen the generous division and for the case that A had chosen the selfish division. We let subjects decide as person B first because we wanted them to be better able to put themselves in the shoes of person B when they decided as person A. Subjects were told in the experimental instructions that at the end of the experiment, it will be randomly determined whether they are person A or B, and their earnings will be calculated based on their actual decisions and the decisions of the subjects they were randomly matched with and who made their decisions in the other role. The corresponding procedure was followed in condition DG3P.

In the second part of the experiment, subjects were randomly assigned to the control condition (TG no info) or, depending on the condition they had been assigned to in the first part, to the corresponding treatment condition (TG info DG2P or TG info DG3P). Subjects were handed condition specific part II instructions on paper, which contained the step-by-step description of the decision situation and an explanation of how subjects' earnings would be calculated based on their decisions and other participants' decisions. After reading the instructions, subjects were asked six control questions about the part II instructions on the screen. After all control questions had been explained to all subjects, the second part of the experiment started.

In part II, subjects first decided as person Y and thereafter they decided as a person X. As person Y, they decided whether to keep the entire amount or return half of it to person X for the case that person X had chosen send. Again, we let subjects decide as person Y first because we wanted them to be better able to empathize with a person Y when deciding as a person X. Subjects decided as person X whether to split or send. In the TG no info condition, they made their decisions unconditionally, whereas in the treatment conditions, they decided whether to split or send in four possible cases: first, person Y had chosen generous as person A; second, person Y had chosen selfish as person A; third, person Y had chosen to accept a selfish division as person B or C; fourth, person Y had chosen to punish a selfish division as person B or C. Subjects' earnings were calculated in the same way as in the first part. No two subjects who had previously been matched in the DG2P or DG3P were paired in the TG, and subjects were told this explicitly in the part II instructions.

At the end of both the first and the second part, we elicited subjects' beliefs regarding other participants' behavior in the corresponding part. We did not incentivise belief elicitation. In the TG no info condition, we asked subjects what they believed, how many out of 100 subjects who participate in this experiment, would choose “keep” and how many would choose “return” as trustees in the TG. In the two treatment conditions, we asked the same question four times, specifying each time what the 100 subjects did in the first part of the experiment. That is, the 100 subjects were characterized as having made a generous division, a selfish division, as having accepted a selfish division, or as having punished a selfish division.

Figs. S1 through S3 in section S1 of the supplementary material (available on the journal's website at www.ehbonline.org) present the experimental instructions as they were presented to subjects in the DG3P (part I) and TG info DG3P condition (part II). Figs. S4 through S9 in section S2 are the shots of the decision and belief elicitation screens subjects saw in the DG3P (part I) and TG info DG3P condition (part II). Instructions as well as decision and belief elicitation screens are translated from German by the authors.

2.4. Data analysis

All test statistics reported in the main article and Figs. 2 through 5 are based on regression model estimations; all corresponding regression tables are presented in the supplementary material (available on the journal's website at www.ehbonline.org). Statistical significance is set at the 5% level (i.e. $\alpha = 0.05$) for two-sided tests. We account for the repeated measures obtained on the same subject by estimating *cluster*-robust standard errors. We use Stata's *margins* command to calculate proportions from logistic regressions and test the statistical significance of the differences between proportions or OLS regression coefficients using Wald tests or F tests of linear hypotheses, respectively.

3. Results

3.1. Generosity and punishment of selfishness

Overall, 44% of subjects chose the generous division when deciding as person A in the dictator game (DG). A smaller proportion of subjects were generous in the DG3P condition (39%) than in the DG2P condition (49%), but the difference is statistically insignificant ($\chi^2_{(1)} = 2.18$, $p = 0.140$). When deciding as person B or person C, a small proportion of subjects punished a generous division (4%), and a considerable proportion of subjects punished a selfish (i.e. self-regarding) division (33%). Subjects in the DG2P condition, when deciding as person B, punished a selfish division in 27% of the cases, whereas subjects in the DG3P condition, when deciding as person C, punished a selfish division in 39% of the cases. This difference too is statistically insignificant ($\chi^2_{(1)} = 2.95$, $p = 0.086$).

Previous studies have found a positive relation between generosity and trustworthiness (Gambetta & Przepiorka, 2014). If, as hypothesized above, punishment is a sign of trustworthiness, generosity and punishment might be positively related as well. Fig. 2 shows the proportion of subjects who chose to punish a selfish division both overall and for the DG2P and DG3P conditions separately. Overall, generosity and punishment are positively related (51% vs. 18%, $\chi^2_{(1)} = 24.09$, $p < 0.001$). Subjects who choose the selfish division are much less likely to punish others for doing the same (18%), whereas subjects who choose the generous division do not appear to have a preference for one or the other. Half the subjects (51%) who choose the generous division in the DG punish others for choosing the selfish division while the other half does not. Also when looking at the DG2P and DG3P conditions separately, generosity and punishment are positively related (39% vs. 15%, $\chi^2_{(1)} = 7.42$, $p = 0.007$ and 67% vs. 21%, $\chi^2_{(1)} = 22.77$, $p < 0.001$). The choice of the selfish division remains a clear sign of subjects' reluctance to punish others (15% and 21%, respectively). Whereas the choice of a generous division is less indicative of subjects' inclination to punish in the DG2P condition (39%), it is a better sign of subjects' inclination to punish in the DG3P condition (67%). This difference is statistically significant (39% vs. 67%, $\chi^2_{(1)} = 6.64$, $p = 0.010$). It is as if generous subjects are less vengeful but more inclined to establish justice on behalf of others.

Let us now turn to one of our central questions, namely whether the inclination (reluctance) to punish a selfish division in the DG is indicative of one's trustworthiness (untrustworthiness).

3.2. Trustworthiness

When deciding as person Y (i.e. trustee in the TG), almost half of the 186 subjects (49%) choose to return half of the CHF 16 they gained from being sent CHF 8 by person X (i.e. truster in the TG). But only if one's inclination to punish a selfish division as person B or C in the DG is positively related with one's inclination to return half the amount as person Y in the TG (our measure of trustworthiness) can punishment be conceived as a sign of trustworthiness.

Fig. 3 shows subjects' trustworthiness contingent on what these subjects decided in the first part of the experiment, overall as well as for the DG2P and DG3P conditions separately. Although overall, punishers tend to be more trustworthy (56%) than non-punishers (46%), this difference is statistically insignificant ($\chi^2_{(1)} = 1.70$, $p = 0.193$). Thus, we find no support for hypothesis H1, that punitive preferences and trustworthiness are positively related. Moreover, we gain no new insights when testing hypothesis H3, that the positive relation between punitive preferences and trustworthiness is stronger in the DG3P condition than in the DG2P condition. Neither in the DG2P nor in the DG3P condition are punishers' and non-punishers' trustworthiness significantly different from each other (48% vs. 46%, $\chi^2_{(1)} = 0.04$, $p = 0.837$ and 61% vs. 46%, $\chi^2_{(1)} = 2.18$, $p = 0.140$).

In accordance with hypothesis H5, we find that overall, generous subjects are significantly more trustworthy than selfish subjects (73% vs. 30%, $\chi^2_{(1)} = 42.45$, $p < 0.001$). This corroborates that generosity can be a reliable sign of trustworthiness and selfishness a reliable sign of untrustworthiness. Moreover, selfishness turns out to be a better sign of untrustworthiness in the DG2P condition than in the DG3P condition (17% vs. 40%, $\chi^2_{(1)} = 7.49$, $p = 0.006$).

Despite the fact that the punishment of an unequal division of money, either on one's own or someone else's part, is not indicative of one's trustworthiness, punishers might still be perceived as being more trustworthy than non-punishers, and subjects might act on these beliefs and correspondingly trust punishers more than non-punishers. Next, we describe how subjects decided as trusters in the TG, when they knew how the trustee had decided in the first part of the experiment, and when they received no information about the trustee. We employ two measures of trust: subjects' beliefs about the trustworthiness of trustees, and subjects' decisions as trusters in the TG. We first look at subjects' beliefs and discuss the results regarding their behavior thereafter.

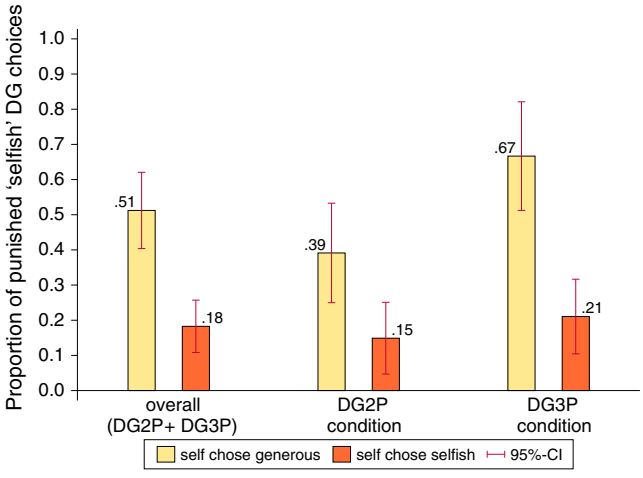

Fig. 2. The figure shows the proportions of subjects' punishment decisions conditional on whether these subjects chose a generous or selfish division as dictators in the dictator game. The overall proportions are based on the data from both the 2nd party punishment and 3rd party punishment condition. The figure is based on the regression model estimations presented in Table S1 in the supplementary material (available on the journal's website at www.ehbonline.org).

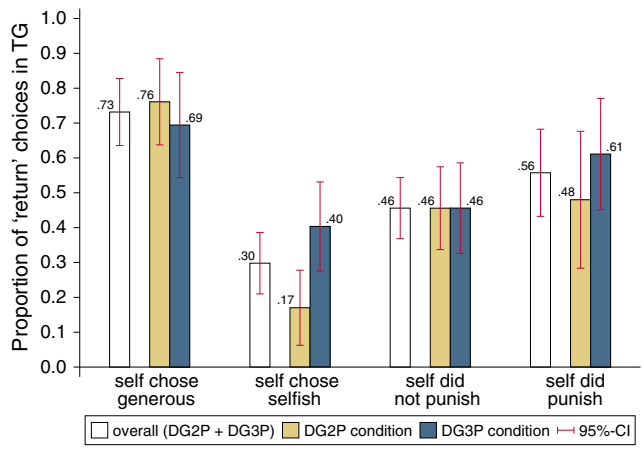

Fig. 3. The figure shows the proportions of subjects' return decisions as trustees in the trust game conditional on whether these subjects chose a generous or selfish division as dictators in the dictator game, and whether they did not punish or punished another subject for choosing a selfish division in the dictator game. The overall proportions are based on the data from both the 2nd party punishment and 3rd party punishment condition. The figure is based on the regression model estimations presented in Table S2 in the supplementary material (available on the journal's website at www.ehbonline.org).

3.3. Trust: beliefs

Fig. 4 shows subjects' average beliefs regarding trustees' trustworthiness (i.e. proportion of return choices in the TG), contingent on what these subjects know about the decisions trustees made in the first part of the experiment. In the no info condition, in which subjects were not given any information about trustees, subjects reported on average that they believed 46% of trustees to be trustworthy. In the treatment conditions, subjects state their beliefs contingent on what the trustee did in the first part of the experiment.

Overall, subjects believe on average 48% of punishers and 49% of non-punishers to be trustworthy; this difference is statistically insignificant ($F_{1,185} = 0.13$, $p = 0.718$). Moreover, subjects' average beliefs regarding the trustworthiness of punishers and non-punishers are statistically indistinguishable from subjects' average beliefs regarding the trustworthiness of trustees about whom they have no additional information (48% vs. 46%, $F_{1,185} = 0.16$, $p = 0.692$ and 49% vs. 46%, $F_{1,185} = 0.54$, $p = 0.463$). From this we conclude that, based on our first measure of trust, we do not find any empirical support for hypothesis H2, that observers infer trustworthiness from punitive acts.

When testing hypothesis H4, that subjects' inference of trustworthiness from punitive acts will be less equivocal in condition DG3P than in condition DG2P, a somewhat unexpected pattern emerges (see Fig. 4). In the DG2P condition, punishers are on average believed to be significantly less trustworthy than non-punishers (40% vs. 53%, $F_{1,185} = 6.60$, $p = 0.011$); and in the DG3P condition, non-punishers are on average believed to be significantly less trustworthy than punishers (45% vs. 55%, $F_{1,185} = 4.80$, $p = 0.030$). However, neither of these average beliefs is significantly different from the corresponding average belief in the no info condition (53% vs. 46%, $F_{1,185} = 1.42$, $p = 0.234$; 40% vs. 46%, $F_{1,185} = 1.12$, $p = 0.291$; 45% vs. 46%, $F_{1,185} = 0.03$, $p = 0.867$; 55% vs. 46%, $F_{1,185} = 2.57$, $p = 0.111$).

Our results clearly support hypothesis H6. Overall, subjects believe on average that 65% of the generous trustees and 31% of the selfish trustees are trustworthy ($F_{1,185} = 134.04$, $p < 0.001$). What is more, subjects believe generous trustees to be significantly more trustworthy ($F_{1,185} = 26.34$, $p < 0.001$), and they believe selfish trustees to be significantly less trustworthy ($F_{1,185} = 15.95$, $p < 0.001$), than trustees about whom they have no additional information (46%).

3.4. Trust: behavior

Fig. 5 shows subjects' trust rates (i.e. proportion of send choices in the TG) contingent on what these subjects know about the decisions trustees made in the first part of the experiment. Without additional information about trustees, the overall average trust rate is 56%. Trust rates tend to differ depending on whether or not the trustee punished a selfish division; we find that punishers are trusted less on average (40%) than non-punishers (53%), but this difference is statistically insignificant ($\chi^2_{(1)} = 3.47$, $p = 0.063$). What is more, in comparison with trustees in the no info condition, about whom no information is available, non-punishers are neither trusted more nor less (53% vs. 56%, $\chi^2_{(1)} = 0.13$, $p = 0.723$), whereas punishers are trusted significantly less (40% vs. 56%, $\chi^2_{(1)} = 4.40$, $p = 0.036$). The empirical evidence thus far still lacks support for hypothesis H2.

A similar pattern as in the previous section emerges if, testing hypothesis H4, we look at the trust rates in punishers and non-punishers in the DG2P and DG3P conditions separately. In the DG2P condition, punishers are trusted significantly less than non-punishers (28% vs. 69%, $\chi^2_{(1)} = 22.88$, $p < 0.001$). In the DG3P condition, non-punishers tend to be trusted less than punishers, but, unlike for beliefs, this difference is statistically insignificant (38% vs. 53%, $\chi^2_{(1)} = 2.37$, $p = 0.124$). A possible explanation for the finding in the DG2P condition is that second-party punishment is predominantly interpreted as vengeful, and vengeful subjects are distrusted more. Note that significantly less trust is placed in seemingly vengeful trustees, than in trustees about whom no additional information is available (28% vs. 54%, $\chi^2_{(1)} = 6.70$, $p = 0.010$). However, trust in non-vengeful trustees is not significantly higher than in the no info condition (69% vs. 54%, $\chi^2_{(1)} = 2.07$, $p = 0.151$).

Consistent with the results obtained in terms of subjects' beliefs, we find that generous trustees are trusted significantly more than selfish trustees (75% vs. 22%, $\chi^2_{(1)} = 85.49$, $p < 0.001$), and these trust rates differ significantly from the rate observed in the no info condition (75% vs. 56%, $\chi^2_{(1)} = 7.61$, $p = 0.006$ and 22% vs. 56%, $\chi^2_{(1)} = 23.82$, $p < 0.001$). Again, this is clear evidence in support of hypothesis H6. Further details on the statistical analysis are provided in the supplementary material (available on the journal's website at www.ehbonline.org), section S3.

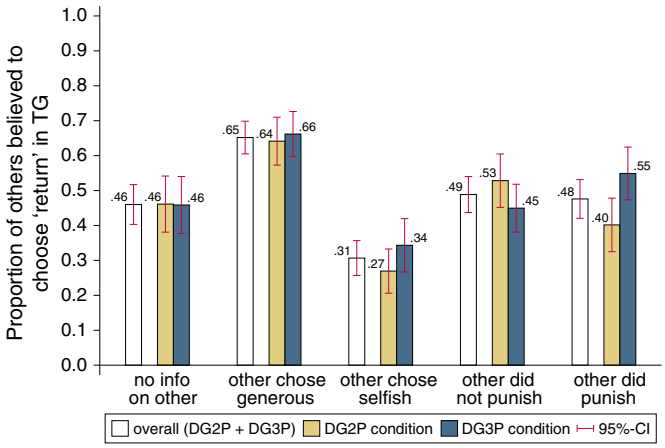

Fig. 4. The figure shows subjects' average beliefs regarding the proportion of trustees' return decisions in the trust game contingent on what these subjects know about the decisions trustees made in the first part of the experiment. In the no info condition, subjects state their beliefs unconditionally, that is, without being given any information on what trustees did in the first part of the experiment. In the other conditions, subjects state their beliefs conditional on whether trustees chose a generous or selfish division as dictators in the dictator game, or whether trustees did not punish or punished another subject for choosing a selfish division in the dictator game. The overall proportions are based on the data from both the 2nd party punishment and 3rd party punishment condition. The figure is based on the regression model estimations presented in Table S3 in the supplementary material (available on the journal's website at www.ehbonline.org).

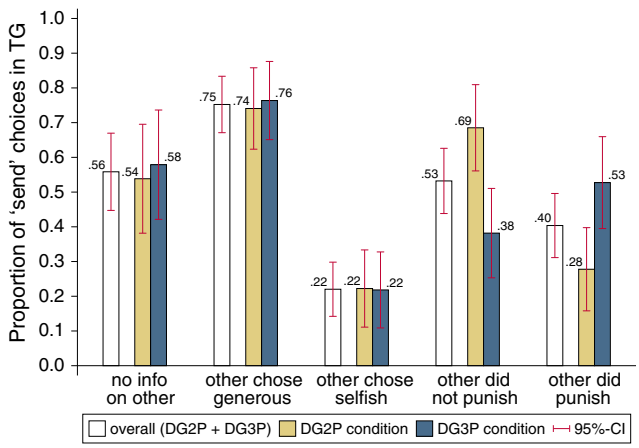

Fig. 5. The figure shows the proportions of subjects' send decisions as trusters in the trust game contingent on what these subjects know about the decisions trustees made in the first part of the experiment. In the no info condition, subjects make their decisions unconditionally, that is, without being given any information on what trustees did in the first part of the experiment. In the other conditions, subjects make their decisions conditional on whether trustees chose a generous or selfish division as dictators in the dictator game, or whether trustees did not punish or punished another subject for choosing a selfish division in the dictator game. The overall proportions are based on the data from both the 2nd party punishment and 3rd party punishment condition. The figure is based on the regression model estimations presented in Table S4 in the supplementary material (available on the journal's website at www.ehbonline.org).

4. Discussion

We address the question whether peer-punishment can function as a type-separating signal of a punisher's cooperative intent. In a laboratory experiment, we test two necessary conditions for a signaling account of peer-punishment to be plausible. First, punishment and cooperative intent must be positively related (hypothesis H1); second, individuals must infer cooperative intent from punishment (H2). Moreover, we distinguish between second-party and third-party punishment because we expect that third-party punishment is a better sign of cooperative intent than second-party punishment, both in terms of a stronger positive relation (H3) and in terms of a less equivocal interpretation by observers (H4). Finally, we compare punitive acts with acts of generosity as signs of trustworthiness. It has been shown theoretically and empirically that generosity and cooperative intent are positively related (H5) and that observers infer cooperative intent from acts of generosity (H6) (Barclay, 2004; Fehrler & Przepiorka, 2013; Gambetta & Przepiorka, 2014; Bliege Bird & Power, 2015). In our experiment, we use the binary dictator game with second-party and third-party punishment to measure subjects' generosity and punitive preferences, and we use the trust game to measure subjects' trust and their cooperative intent in terms of their trustworthiness.

We find that generous subjects are more trustworthy and are trusted more than selfish (i.e. self-regarding) subjects or subjects about whom no information is available. However, with regard to punitive acts, we do not find any evidence in support of our hypotheses, neither in the second-party nor in the third-party punishment condition. That is, punishment and trustworthiness are not positively related and individuals do not trust punishers more than non-punishers. To the contrary, we find that second-party punishers are trusted less than their non-punishing counterparts and less than trustees about whom no information is available. This finding is consistent with the idea that second-party punishment is interpreted as vengeful and vengefulness perceived as an untrustworthy-making property. However, in our experiment, seemingly vengeful subjects are not less trustworthy than seemingly non-vengeful subjects (see Fig. 3).

The enforcement of cooperation through peer-punishment is essential for social cohesion and the functioning of societies. Despite its central role in establishing and maintaining social order, the ultimate reasons for why individuals engage in peer-punishment are not fully understood. At the proximate level of analysis, it has been suggested that punitive preferences trigger the punishment of deviant peers in one-time-only encounters. To determine the possible function punitive preferences evolved to fulfill, recent scholarship has aimed at identifying the motives punitive acts comprise. The evidence thus far suggests that different motives may be at work when individuals decide to punish their deviant peers, and these motives may vary across the contexts in which peer-punishment occurs (Raihani & Bshary, 2015a). The range of findings has provided a breeding ground for new speculations about the evolution of punitive preferences. Here we take a somewhat different approach. We start from a mechanism which may have facilitated the evolution of punitive preferences, derive the conditions under which such an ultimate explanation is plausible, and test these conditions in a controlled experiment (Krasnow et al., 2012).

It has been argued that pro-social acts in general and peer-punishment in particular could function as a type-separating signal of an individual's cooperative intent (Gintis et al., 2001). That is, if cooperators are more likely to punish defectors than defectors are to punish defectors, then punishment will be a reliable sign of an individual's cooperative intent. Since cooperative individuals are more attractive partners for mutually beneficial social exchange than defectors are, punishment as a sign of cooperative intent should spread through positive assortment and partner choice (Eshel & Cavalli-Sforza, 1982; Baumard, André, & Sperber, 2013). The benefits which accrue in social exchange will partly compensate punishers for the costs they incur when punishing defectors, and may ease the conditions under which punitive preferences can evolve (Fehrler & Przepiorka, 2013).

Our results seem to preclude that the punishment of a selfish division of money can be a sign of trustworthiness. However, our results do not preclude that other types of punitive acts convey trustworthiness, or that the punishment of selfishness conveys other types of information. With regard to the former, it has been shown that moral judgments of unfair behavior can be a sign of trustworthiness in social exchange (Simpson et al., 2013). With regard to the latter, it seems plausible that, by punishing unfair behavior, one can build a reputation for being a punisher and deter future attempts of exploitation (Barclay, 2006; Nelissen, 2008; dos Santos et al., 2011; Gordon et al., 2014). However, reputation building is restricted to relatively small groups, where punishers and non-punishers can be identified and targeted by reward or (second-order) punishment, respectively. In large societies, in which anonymous interactions are frequent, it is more decisive what one signals about oneself to strangers (Marlowe et al., 2011), because such signals are relevant for partner choice rather than for the maintaining of existing relationships (Delton et al., 2011; Raihani & Bshary, 2015a). It has been shown that, if given the chance, third parties might prefer to be generous and compensate the victim of a selfish act rather than punish the perpetrator (Chavez & Bicchieri, 2013). Our evidence shows that generosity is indeed a better sign of a potential partner's trustworthiness than the punishment of selfishness.

Supplementary materials

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.evolhumbehav.2015.12.003.

Acknowledgements

We would like to thank Maria Bigoni, Stefania Bortolotti, Vincent Buskens, Marco Casari, Andreas Diekmann, Sebastian Fehrler, Diego Gambetta, Donna Harris, Erik Mohlin, Ryan Murphy, Werner Raub, Manuela Vieth, Erte Xiao and the handling editor, Debra Lieberman, for their perceptive comments and suggestions, and Nadja Jehli for excellent research assistance. W.P. gratefully acknowledges the financial support of the OUP John Fell Fund [grant no. 103/848].

References

Axelrod, R. (1986). An evolutionary approach to norms. *American Political Science Review*, *80*, 1095–1111.
Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 15924–15927.
Barclay, P. (2004). Trustworthiness and competitive altruism can also solve the "tragedy of the commons". *Evolution and Human Behavior*, *25*, 209–220.
Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior*, *27*, 325–344.
Barclay, P. (2012). Proximate and ultimate causes of punishment and strong reciprocity. *Behavioral and Brain Sciences*, *35*, 16–17.
Baumard, N. (2010). Has punishment played a role in the evolution of cooperation? A critical review. *Mind & Society*, *9*, 171–192.
Baumard, N., André, J. -B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, *36*, 59–78.
Becker, G. S. (1976). Altruism, egoism and genetic fitness: Economics and sociobiology. *Journal of Economic Literature*, *14*, 817–826.
Bliege Bird, R., & Power, E. A. (2015). Prosocial signaling and cooperation among Martu hunters. *Evolution and Human Behavior*, *36*, 389–397.
Bliege Bird, R., & Smith, E. A. (2005). Signaling theory, strategic interaction, and symbolic capital. *Current Anthropology*, *46*, 221–248.
Bone, J. E., & Raihani, N. J. (2015). Human punishment is motivated by both a desire for revenge and a desire for equality. *Evolution and Human Behavior*, *36*, 323–330.
Boyd, R., Gintis, H., & Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science*, *328*, 617–620.
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*, 3531–3535.
Boyd, R., & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology*, *13*, 171–195.

Brañas-Garza, P., Espín, A. M., Exadaktylos, F., & Herrmann, B. (2014). Fair and unfair punishers coexist in the Ultimatum Game. *Scientific Reports*, *4*, 6025.
Brandts, J., & Charness, G. (2011). The strategy versus the direct-response method: A first survey of experimental comparisons. *Experimental Economics*, *14*, 375–398.
Carpenter, J., & Matthews, P. H. (2009). What norms trigger punishment? *Experimental Economics*, *12*, 272–288.
Chavez, A. K., & Bicchieri, C. (2013). Third-party sanctioning and compensation behavior: Findings from the ultimatum game. *Journal of Economic Psychology*, *39*, 268–277.
Coleman, J. S. (1990). *Foundations of social theory*. Cambridge (MA): The Belknap Press of Harvard University Press.
Dasgupta, P. (1988). Trust as a commodity. In D. Gambetta (Ed.), *Trust: Making and breaking cooperative relations* (pp. 49–72). Oxford: Basil Blackwell.
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*, 1254–1258.
Delton, A. W., Krasnow, M. M., Cosmides, L., & Tooby, J. (2011). Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 13335–13340.
Diekmann, A., Jann, B., Przepiorka, W., & Wehrli, S. (2014). Reputation formation and the evolution of cooperation in anonymous online markets. *American Sociological Review*, *79*, 65–85.
Diekmann, A., & Przepiorka, W. (2015). Punitive preferences, monetary incentives and tacit coordination in the punishment of defectors promote cooperation in humans. *Scientific Reports*, *5*, 10321.
dos Santos, M., Rankin, D. J., & Wedekind, C. (2011). The evolution of punishment through reputation. *Proceedings of the Royal Society B*, *278*, 371–377.
Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don't punish. *Nature*, *452*, 348–351.
Engelmann, D., & Fischbacher, U. (2009). Indirect reciprocity and strategic reputation building in an experimental helping game. *Games and Economic Behavior*, *67*, 399–407.
Eshel, I., & Cavalli-Sforza, L. L. (1982). Assortment of encounters and evolution of cooperativeness. *Proceedings of the National Academy of Sciences of the United States of America*, *79*, 1331–1335.
Fehr, E. (2009). On the economics and biology of trust. *Journal of the European Economic Association*, *7*, 235–266.
Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*, 63–87.
Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature*, *13*, 1–25.
Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*, 137–140.
Fehrler, S., & Przepiorka, W. (2013). Charitable giving as a signal of trustworthiness: Disentangling the signaling benefits of altruistic acts. *Evolution and Human Behavior*, *34*, 139–145.
FeldmanHall, O., Sokol-Hessner, P., Bavel, J. J. V., & Phelps, E. A. (2014). Fairness violations elicit greater punishment on behalf of another than for oneself. *Nature Communications*, *5*, 5306.
Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*, 171–178.
Fowler, J. H. (2005). Altruistic punishment and the origin of cooperation. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 7047–7049.
Gächter, S., Renner, E., & Sefton, M. (2008). The long-run benefits of punishment. *Science*, *322*, 1510.
Gambetta, D. (2009). Signaling. In P. Hedström, & P. Bearman (Eds.), *The Oxford handbook of analytical sociology* (pp. 168–194). Oxford: Oxford University Press.
Gambetta, D., & Hamill, H. (2005). *Streetwise: How taxi drivers establish their customers' trustworthiness*. New York: Russell Sage Foundation.
Gambetta, D., & Przepiorka, W. (2014). Natural and strategic generosity as signals of trustworthiness. *PloS One*, *9*, e97533.
Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology*, *206*, 169–179.
Gintis, H., Smith, E. A., & Bowles, S. (2001). Costly signaling and cooperation. *Journal of Theoretical Biology*, *213*, 103–119.
Gordon, D. S., Madden, J. R., & Lea, S. E. G. (2014). Both loved and feared: Third party punishers are viewed as formidable and likeable, but these reputational benefits may only be open to dominant individuals. *PloS One*, *9*, e110045.
Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences*, *35*, 1–15.
Gummerum, M., & Chu, M. T. (2014). Outcomes and intentions in children's, adolescents', and adults' second- and third-party punishment behavior. *Cognition*, *133*, 97–103.
Gürerk, Ö., Irlenbusch, B., & Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science*, *312*, 108–111.
Hamilton, W. D. (1963). The evolution of altruistic behavior. *American Naturalist*, *97*, 354–356.
Hardin, R. (2002). *Trust and trustworthiness*. New York: Russell Sage Foundation.
Harris, D., Herrmann, B., Kontoleon, A., & Newton, J. (2015). Is it a norm to favour your own group? *Experimental Economics*, *18*, 491–521.
Hauert, C., Traulsen, A., Brandt, H., Nowak, M. A., & Sigmund, K. (2007). Via freedom to coercion: The emergence of costly punishment. *Science*, *316*, 1905–1907.
Helbing, D., Szolnoki, A., Perc, M., & Szabó, G. (2010). Evolutionary establishment of moral and double moral standards through spatial interactions. *PLoS Computational Biology*, *6*, e1000758.
Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., ... Ziker, J. (2006). Costly punishment across human societies. *Science*, *312*, 1767–1770.
Jordan, J. J., McAuliffe, K., & Rand, D. G. (2015). The effects of endowment size and strategy method on third party punishment. *Experimental Economics* (forthcoming).
Krasnow, M. M., Cosmides, L., Pedersen, E. J., & Tooby, J. (2012). What are punishment and reputation for? *PloS One*, *7*, e45662.
Kreps, D. (1990). Corporate culture and economic theory. In J. E. Alt, & K. A. Shepsle (Eds.), *Perspectives on positive political economy* (pp. 90–143). Cambridge (MA): Cambridge University Press.
Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior*, *28*, 75–84.
Leibbrandt, A., & López-Pérez, R. (2012). An exploration of third and second party punishment in ten simple games. *Journal of Economic Behavior & Organization*, *84*, 753–766.
Marlowe, F. W., Berbesque, J. C., Barrett, C., Bolyanatz, A., Gurven, M., & Tracer, D. (2011). The 'spiteful' origins of human cooperation. *Proceedings of the Royal Society B*, *278*, 2159–2164.
Nelissen, R. M. A. (2008). The price you pay: cost-dependent reputation effects of altruistic punishment. *Evolution and Human Behavior*, *29*, 242–248.
Oliver, P. (1980). Rewards and punishments as selective incentives for collective action: Theoretical investigations. *American Journal of Sociology*, *85*, 1356–1375.
Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a sword: Self-governance is possible. *American Political Science Review*, *86*, 404–417.
Pedersen, E. J., Kurzban, R., & McCullough, M. E. (2013). Do humans really punish altruistically? A closer look. *Proceedings of the Royal Society B*, *280*, 20122723.
Przepiorka, W., & Diekmann, A. (2013). Individual heterogeneity and costly punishment: A volunteer's dilemma. *Proceedings of the Royal Society B*, *280*, 20130247.
Raihani, N. J., & Bshary, R. (2011). The evolution of punishment in n-player public goods games: A volunteer's dilemma. *Evolution*, *65*, 2725–2728.
Raihani, N. J., & Bshary, R. (2015a). The reputation of punishers. *Trends in Ecology and Evolution*, *30*, 98–103.
Raihani, N. J., & Bshary, R. (2015b). Third-party punishers are rewarded, but third-party helpers even more so. *Evolution*, *69*, 993–1003.
Roberts, G. (2013). When punishment pays. *PloS One*, *8*, e57378.
Rockenbach, B., & Milinski, M. (2011). To qualify as a social partner, humans hide severe punishment, although their observed cooperativeness is decisive. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 18307–18312.
Selten, R. (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. In H. Sauermann (Ed.), *Beiträge zur experimentellen Wirtschaftsforschung* (pp. 136–168). Tübingen: Mohr Siebeck.
Sigmund, K. (2007). Punish or perish? Retaliation and collaboration among humans. *Trends in Ecology and Evolution*, *22*, 593–600.
Simpson, B., Harrell, A., & Willer, R. (2013). Hidden paths from morality to cooperation: Moral judgments promote trust and trustworthiness. *Social Forces*, *91*, 1529–1548.
Spence, M. A. (1974). *Market signaling: Informational transfer in hiring and related screening processes*. Cambridge (MA): Harvard University Press.
Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, *46*, 35–57.
West, S. A., El Mouden, C., & Gardner, A. (2011). Sixteen common misconceptions about the evolution of cooperation in humans. *Evolution and Human Behavior*, *32*, 231–262.
Willer, R., Kuwabara, K., & Macy, M. W. (2009). The false enforcement of unpopular norms. *American Journal of Sociology*, *115*, 451–490.
Xiao, E., & Houser, D. (2005). Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 7398–7401.
Zahavi, A. (1975). Mate selection—a selection for a handicap. *Journal of Theoretical Biology*, *53*, 205–214.
Zahavi, A. (1977). The cost of honesty: Further Remarks on the handicap principle. *Journal of Theoretical Biology*, *67*, 603–605.
Zahavi, A. (1995). Altruism as a handicap: The limitation of kin selection and reciprocity. *Journal of Avian Biology*, *26*, 1–3.
Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, *13*, 75–98.