frontiers in
ECOLOGY AND EVOLUTION

ORIGINAL RESEARCH ARTICLE
published: 01 August 2014
doi: 10.3389/fevo.2014.00041

Exposure to superfluous information reduces cooperation and increases antisocial punishment in reputation-based interactions

Miguel dos Santos*[1], *Victoria A. Braithwaite*[2,3] *and Claus Wedekind*[1,2]

[1] *Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland*
[2] *Institute of Cell, Animal, and Population Biology, University of Edinburgh, Edinburgh, Scotland*
[3] *Department of Ecosystem Science and Management, Center for Brain, Behavior and Cognition, Pennsylvania State University, University Park, PA, USA*

Edited by:
Sasha Raoul Xola Dall, University of Exeter, UK

Reviewed by:
Julie Morand-Ferron, University of Ottawa, Canada
Pete C. Trimmer, University of Bristol, UK
Peter L. Hurd, University of Alberta, Canada

***Correspondence:**
Claus Wedekind, Department of Ecology and Evolution, University of Lausanne, Biophore, CH-1015 Lausanne, Switzerland
e-mail: claus.wedekind@unil.ch

Human cooperation is often based on reputation gained from previous interactions with third parties. Such reputation can be built on generous or punitive actions, and both, one's own reputation and the reputation of others have been shown to influence decision making in experimental games that control for confounding variables. Here we test how reputation-based cooperation and punishment react to a disruption of the cognitive processing in different kinds of helping games with observers. Saying a few superfluous words before each interaction was used to possibly interfere with working memory. In a first set of experiments, where reputation could only be based on generosity, the disturbance reduced the frequency of cooperation and lowered mean final payoffs. In a second set of experiments where reputation could only be based on punishment, disturbance increased the frequency of antisocial punishment (i.e., of punishing those who helped) and reduced the frequency of punishing defectors. Our findings suggest that working memory can easily be constraining in reputation-based interactions within experimental games, even if these games are based on a few simple rules with a visual display that provides all the information the subjects need to play the strategies predicted from current theory. Our findings also highlight a weakness of experimental games, namely that they can be very sensitive to environmental variation and that quantitative conclusions about antisocial punishment or other behavioral strategies can easily be misleading.

Keywords: indirect reciprocity, game theory, experimental games, image score, punishment reputation, sanctions, cognitive constraints, helping behavior

INTRODUCTION

Explaining the evolution of cooperative behaviors in humans is a longstanding challenge for economists, social scientists, and evolutionary biologists. A variety of mechanisms have been identified that promote cooperation and that require information about the behavior of other individuals (Brosnan et al., 2010; Bshary and Bronstein, 2011). Information about one's cooperative strategies can be obtained from direct interactions (Trivers, 1971; Axelrod and Hamilton, 1981; Sigmund, 2010) or be inferred from interactions among others in the population (Alexander, 1987; Nowak and Sigmund, 2005; Earley, 2010; Sigmund, 2012). Human cooperation is often built on reputation that reflects previous behavioral decisions and that may be based on own observations or on gossip (Sommerfeld et al., 2007). A reputation of being generous, for example, can build up trust and thereby increase cooperation frequency (Nowak and Sigmund, 1998; Wedekind and Milinski, 2000; Wedekind and Braithwaite, 2002; Yoeli et al., 2013), whereas a non-generous reputation can lead to social exclusion (Guala, 2012; Sasaki and Uchida, 2013) and other types of punishment (Sigmund, 2007; Raihani et al., 2012). Apart from reputation based on generosity, there are other kinds of reputational effects that influence cooperation in humans. A reputation based on punishment, for example, may have played a key role in the evolution of punishment that promotes cooperation within groups (Brandt et al., 2003; Gardner and West, 2004; Hilbe and Sigmund, 2010; dos Santos et al., 2011, 2013). Various kinds of reputation may therefore affect partner choice (Fu et al., 2008; Sylwester and Roberts, 2010) and may thereby create a biological market (Barclay, 2013).

Reputation games usually require the translation of observed behavioral decisions (or of gossip) into an image score that then needs to be continuously updated for all potential partners within social groups. The success of cooperation through direct or indirect reciprocity is therefore expected to rely heavily on the cognitive abilities of the actors (Stevens and Hauser, 2004; Stevens et al., 2005; Brosnan et al., 2010). Working memory (Becker and Morris, 1999) and partner recognition seem to be crucial for processing past actions, the reputation of social partners, and for computing the adequate response to a social dilemma (Stevens et al., 2005, 2011; Volstorf et al., 2011; Moreira et al., 2013). Any alteration of these capacities might influence how cooperation is achieved or whether it is achieved at all (Horvath et al., 2012).

For example, cooperation through direct reciprocity dropped in zebra finches (*Taenipygia guttata*) when the birds' cognitive ability was experimentally reduced (Larose and Dubois, 2011), and humans playing a repeated Prisoner's Dilemma game switched their strategies to less memory-demanding ones (Pavlovian to Tit-for-Tat like strategies) when their working memory was experimentally constrained (Milinski and Wedekind, 1998).

Theory predicts that very simple reputation-based strategies can lead to cooperation in games that are far less complex and memory demanding than real-life situations. For example, a score that reveals the number of times a social partner helped or punished others in previous interactions would be sufficient to trigger cooperation in simple repeated helping games based on indirect reciprocity (e.g., Nowak and Sigmund, 1998; dos Santos et al., 2011). Experiments that were specifically designed to test these models found the kind of reputation formation and the increased cooperation that were predicted (Wedekind and Milinski, 2000; Wedekind and Braithwaite, 2002; dos Santos et al., 2013). These experiments seemed to require comparatively little memory because all the information that theory predicted to be relevant was provided in simple graphical form on a screen (e.g., a moveable arrow that indicated a level on a scale). It is, however, still unclear whether subjects in such experimental games use the simple rules that are predicted by theory (e.g., integrating the information provided on the display) or whether they use different and potentially more memory-demanding strategies that might lead to outcomes that are similar to those predicted from the simpler models. Indeed, the typical assumptions about human cognition that are made in current game theory seem unrealistic (Stevens et al., 2011). The associative nature of human memory contrasts significantly with the type of memory that is often implemented in game-theoretical models (Stevens et al., 2011; Volstorf et al., 2011; Bell et al., 2013). A distinctive feature between the simple memory strategies implemented in most reputation-based models and the potentially more memory-demanding strategies played by humans may not be whether cooperation is achieved, but how sensitive behavioral strategies within a given type of social interactions are to disturbance. We therefore tested whether reputation games are sensitive to superfluous verbal information (similar treatments have been shown to affect working memory of people with or without dementia, see Rouleau and Belleville, 1996; Belleville et al., 2003). We introduced this treatment into two types of reputation games, one where reputation can only be based on generosity and one where reputation can only be based on punishment. We based our analyses on experimental protocols that were successfully used before to test predictions derived from game theory (Wedekind and Braithwaite, 2002; dos Santos et al., 2013) and that can both lead to reputation-based cooperation with simple behavioral rules using the simple graphical information provided on display only (consistent with the recent theory), or with more sophisticated and potentially more memory-demanding strategies.

METHODS

EXPERIMENTAL SET-UP

A total of 185 university students participated in 22 separate groups. To play anonymously within groups, players were asked to choose a plug from an impenetrable tangle of cables, connect it to a box, and chose one of 10 isolated cubicles in juxtaposition from where each player could see the same projector screen that displayed the details of the game. To reveal a choice, players could secretly push one of two buttons inside the box (Wedekind and Milinski, 2000; Wedekind and Braithwaite, 2002). The buttons were connected via the tangle of cables and a switchboard to a green and a red light, respectively, as in Wedekind and Braithwaite (2002). One of two procedures was used to ensure full anonymity of each player within a group and toward the experimenters. In study 1 (see below), players learned their own identification number by drawing one of many sheets with individual sequences of four colors (red and/or green) from a pot, reading it in secret, and then, after all players had drawn a sheet each, pushing their individual color code after the operator had announced an ID number and switched the respective connection on. Each player realized their player ID when they saw their code sequence flashed out by the light display. In study 2, small bulbs inside the box indicated to a player when he/she was connected via the switchboard and hence a choice was due.

The experimenter read the game instructions (available from the authors on request). Each player received an initial endowment that could be used in the game. No information was given about the total number of interactions that would be played. At the end of each game, players were paid out in a way that retained their anonymity: the individual gains were put in envelopes marked with the player IDs (numbers from 1 to 10) and distributed on a table in a room. Each student was sent alone into the room to take the money from the envelope marked with his or her player ID, put the envelop back on the table, and leave the room so that the next student could be sent in. A guard in the back of the room made sure that no student would touch more than one envelope, e.g., that students could not find out which envelope had already been emptied. Later, all students confirmed with their signature that they had received the amount they expected.

HELPING GAME WITH REPUTATION BASED ON GENEROSITY (STUDY 1)

Seventy-four students of the University of Edinburgh (UK) were distributed among 8 groups of 9 or 10. The students played the same game that 12 other groups of students had played in Wedekind and Braithwaite (2002), i.e., after a practice session (a simultaneous two-player Prisoner's Dilemma) they played a pairwise indirect reciprocity game for real money (initial account of £3.00) and with a new ID. The player in the Donor role could give something to the Receiver (green light) or not (red light). The cost of giving was £0.50, the benefit of receiving £1.00 (the experimenter donated the difference). After this single interaction, a new pair of players was chosen. The players were told that the same pair would never play in the reversed role, i.e., direct reciprocity was not possible. Generosity was recorded as the image score that was suggested by Nowak and Sigmund (1998): giving something increased the image score of the Donor by one unit, not giving decreased it by one unit. This image score was graphically displayed with an arrow that wandered from an initial image score of 0 to a minimum of −6 or a maximum of 6. Both player's histories of giving or not could therefore be displayed with these arrows before each interaction. We played 24 rounds per group.

Each player played once per round as Donor and twice per two rounds as Receiver (e.g., once in round 1 and 2 each, or only twice in round 2).

HELPING GAME WITH REPUTATION BASED ON PUNISHMENT (STUDY 2)

In this study, 111 students of the University of Lausanne (Switzerland) were distributed over 14 groups of 6 to 9. The same equipment and procedures as in study 1 were used to play the same indirect reciprocity game as in study 1, with the following changes: (i) the practice session was waived, (ii) a donation cost 1 CHF for the Receiver to gain 2 CHF (i.e., a change of currency was necessary, but the costs and benefits were similar to the ones in the first study), (iii) after each decision of the Donor, the Receiver had an opportunity to punish the Donor (punishment cost 1 CHF to the punisher and reduced the punished account by 2 CHF), (iv) initial account was 20 CHF to avoid negative accounts, and (v) only the Receiver's ID and his/her punishment reputation was displayed in order to experimentally separate punishment reputation from a scoring of generosity. This punishment reputation was analogous to the image score in the first study, i.e., a score (indicated with an arrow starting at 0, minimum $= -5$, maximum $= +5$) that increased by one unit after the Receiver punished a Donor who had not donated, and decreased by one unit after Receiver had not punished a Donor who had not donated. Each player played 16 times in each of the two roles, i.e., 16 rounds were played in total. In this study, the control groups (see below) were also used as the controls for another study on human discrimination between punishers and non-punishers (dos Santos et al., 2013).

TESTING THE EFFECTS OF CONSTRAINED WORKING MEMORY

Groups were assigned either to a control treatment where they could make decisions in a silent environment (as in Wedekind and Braithwaite, 2002 and in dos Santos et al., 2013) or a disturbance treatment where the experimenters said superfluous words before each decision was due. These phrases were variations of, for example, "Donor number 7, please make your choice," or "Receiver number 2, you are now connected" (translated from the French that was used in the 2nd study). Because players could know that their decision was due by merely looking at the screen or because the small bulb inside their box lit up, we considered that the superfluous talking would act as a distractor and hence impair working memory. The disturbance treatment was applied to 3 of the 8 groups in study 1 and 3 of the 14 groups in study 2.

STATISTICAL ANALYSES

We used generalized linear models (GLM) with binomial distribution to analyze the proportion of cooperative (or punitive) decisions per group. Linear mixed models (LMM) were used to analyze continuous response variables at the individual level, with Group ID as a random factor. Generalized linear mixed models (GLMM) with binomial distribution were used to analyze the Donor's decisions as a function of the Receiver's reputation (image score or punishment score, respectively), with Donor (nested in Group) and Group ID as random factors. Both reputation scores were corrected for group and time effects by subtracting the current group mean score. The statistical analyses were carried out with R 2.10.1 (R Development Core Team, 2010) with the lme4 package for linear and logistic mixed-effect model analyses (Bates and Sarkar, 2007). *P*-values are directed and indicated as P_{dir} for analyses that replicate analogous ones of Wedekind and Braithwaite (2002). We use directed testing in such cases because it reduces both, type I errors for findings that contradict previous ones and type II errors for findings that confirm previous ones, while avoiding the inflation of the alpha value that comes with one-tailed testing (Rice and Gaines, 1994). All other *p*-values are two-tailed.

ETHICAL NOTE

The experiments conformed to the relevant regulatory standards. Participation was voluntary and not linked to any types of course credits. Before the experiments, all participants were informed that the study would be about an evolutionary and economic problem, that they would play for real money in a game that would last for about 1–2 h, that they would be provided with a starting account, that they could keep what would be left of their starting account plus their earnings during the game, that they would play fully anonymously (i.e., anonymous to their colleagues and to us), and that their anonymity would be maintained after the game. The students were recruited on campus with permission from the respective schools. The monetary issues of the experimental games were reviewed and approved by the financial departments of the involved universities (Edinburgh and Lausanne). The experiments that included the option of punishment were also reviewed by the *Commission cantonale (VD) d'éthique de la recherche clinique Sous-Commission III*.

RESULTS

REPUTATION BASED ON GENEROSITY

We found significantly less cooperation in the disturbed groups (i.e., in those that were exposed to the superfluous words) than in the controls (GLM, $z = -4.19$, $P < 0.0001$; **Figure 1A**). Consequently, image scores were on average higher in the control than in the disturbed groups (LMM, $t = -2.76$, $P = 0.007$; **Figure 1B**). Players in control groups finished with higher payoffs (LMM, $t = -1.82$, $P_{\text{dir}} = 0.045$; **Figure 1C**).

Treatment also affected the use of reputation. In the controls, Donors used the Receivers' score to decide whether to donate or not (GLMM, Receivers' score: $z = 5.13$, $P_{\text{dir}} < 0.0001$; **Figure 2A**), as well as their own score (GLMM, Donors' score: $z = -8.1$, $P_{\text{dir}} < 0.0001$; **Figure 2A**). In the disturbed groups, players still used the Receivers' reputation (GLMM, Receivers' score: $z = 5.14$, $P < 0.0001$; **Figure 2B**), but the Donors' own image score had significantly less effect on the Donors' decision to donate (GLMM, $z = 4.32$, $P < 0.0001$; **Figure 2B**), and was only close to significance (Donors' score: $z = -1.95$, $P = 0.051$). Cooperative players received significantly less monetary benefits in the disturbed than in the control groups (**Table 1**; **Figure 3A**).

As expected, the costs of helping led to negative correlations between image score and account during the first rounds of the game in both treatment groups (**Figure 3B**). The Donors' tendency to reward high image scores compensated for the costs

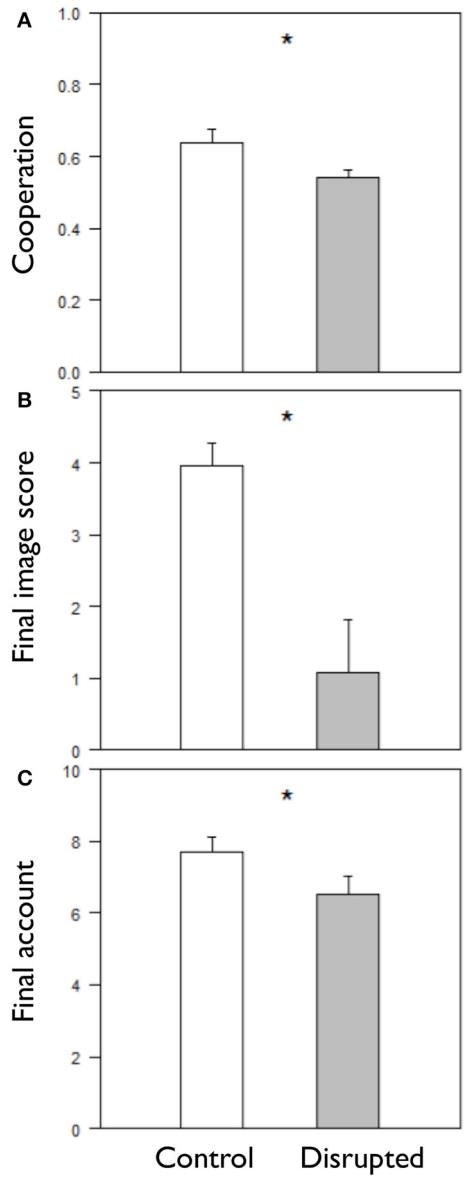

FIGURE 1 | Cooperation in the helping game without punishment (study 1). The figure shows the mean (± SE) per treatment of **(A)** cooperation frequency, **(B)** final image score, and **(C)** final account. See text for statistics (*$p < 0.05$).

of helping after some rounds in the control groups. In the disturbed groups, however, the correlations between image score and account remained negative throughout the 24 rounds, i.e., no significant compensatory effects of reputation could be observed (**Figure 3B**).

REPUTATION BASED ON PUNISHMENT

The experimental disturbance did not significantly affect cooperativeness in the indirect reciprocity games with punishment (GLM, $z = -1.02$, $P = 0.31$; **Figure 4A**). However, disturbance reduced the rate of punishing defection (GLM, $z = -2.02$, $P = 0.043$; **Figure 4B**) and increased the rate of punishing cooperative moves (GLM, $z = 4.69$, $P < 0.0001$; **Figure 4C**). Consequently, overall punishment (either punishing defection or cooperation) did not differ between treatments (GLM, $z = 1.64$, $P = 0.10$). There was no significant treatment effect on mean final accounts (LMM, $t = -0.59$, $P = 0.56$).

Donors used the Receivers' punishment score in the control groups (GLMM, $z = 2.69$, $P = 0.007$; **Figure 5**), but not in the disturbed ones (GLMM, $z = -0.33$, $P = 0.74$; **Figure 5**). The reputational effects of punishment would therefore be expected to compensate at least partly for the costs of punishment. Indeed, the correlation between account and mean punishment score was not significantly negative in the control groups (LMM, $t = -0.68$, $P = 0.49$; see also dos Santos et al., 2013). However, the correlation between account and mean punishment score was not significant in the disturbed groups either (LMM, $t = -1.23$, $P = 0.23$).

DISCUSSION

We found in our first experiment that saying superfluous words to students who play an indirect reciprocity game significantly altered their use of reputation and reduced overall cooperation. When undisturbed, individuals took information about their partners' reputation into account and showed a strong tendency to help those who have helped others before. Their decisions were also influenced by their own reputation, apparently with increasing importance the further their own reputation deviated from the group mean. Therefore, subjects with very low image scores frequently donated even to low-reputation partners. Over time, the tendency of others to reward high image scores, i.e., the long-term benefits of generosity clearly compensated for the immediate costs of building up an image score. Therefore, a necessary condition for the evolution of reputation-based generosity was met (Nowak and Sigmund, 1998). By the end of our game, the tendency of others to reward high image score even resulted in a positive correlation between image score and total income. All these findings confirm previous observations that used the same protocol on other groups of students (Wedekind and Braithwaite, 2002). In the disturbed conditions, however, subjects did not seem to care as much about their own reputation as they would in the undisturbed conditions. It happened comparatively often, for example, that they defected with low reputation partners even though their own score was also low. As a result, average image score, overall cooperativeness, and final payoff were all lower in the disturbed conditions as compared to the controls. Moreover, the immediate costs of generosity did not get compensated during the 24 rounds in our disturbed environment, i.e., the correlation between image score and total income remained negative, and no long-term benefits of generosity could be observed. Therefore, a key condition for the evolution of reputation-based generosity appeared to be not met under the disturbed and possibly memory-constrained conditions.

Our second study controlled experimentally for any possible effects of reputation based on generosity. If reputation was built up in the second experiment, it could only be linked to punishment behavior. dos Santos et al. (2013) recently found that, under such conditions, humans provide more help to those who have a high punishment reputation than those with a low

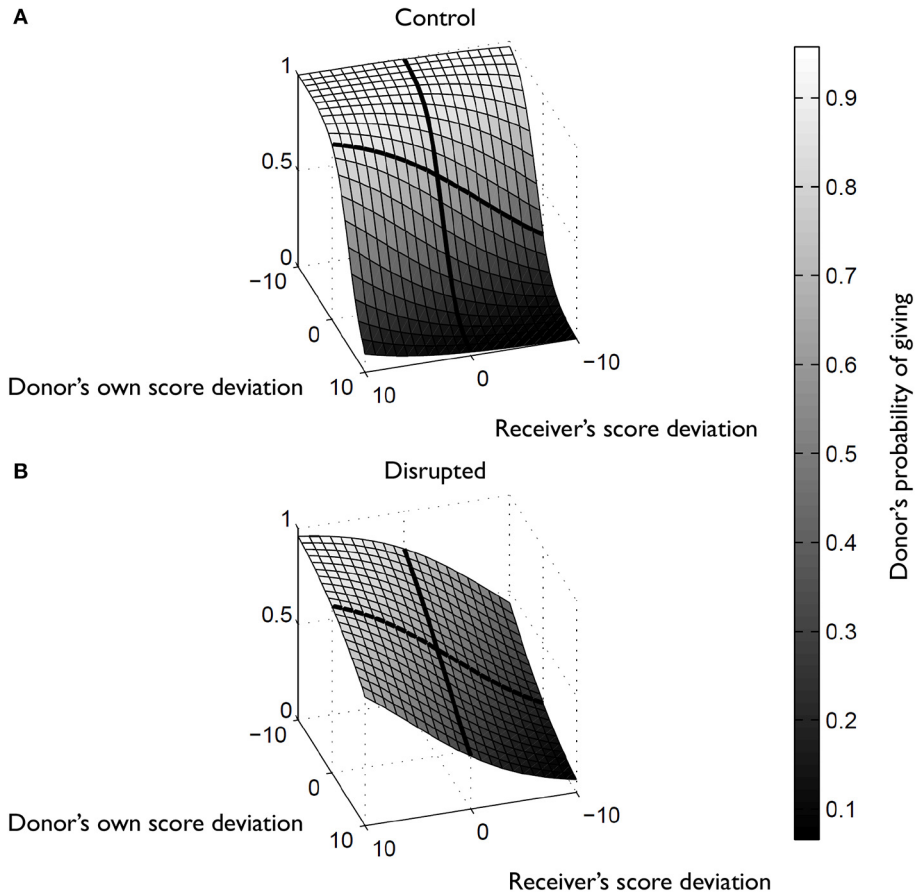

FIGURE 2 | Reputation use study 1. (A) In the control groups, Donors were increasingly more likely to donate as the Receivers' score increased. The Donors were also increasingly likely to donate with decreasing own score, and less likely to donate with increasing own score. **(B)** In disturbed groups, Donors were still more likely to donate as the Receivers' score increased, but effects of Donor's own score seemed less important as in the control condition. The bold lines show the projection of the x and y axes. Scores are the deviation from the group mean per interaction. See text for statistics.

Table 1 | Linear mixed-effect model fit on the individual final account explained by mean image score in study 1.

	Estimate (SE)	*t*	*P*
CONTROLS			
Intercept	6.71 (0.64)	10.38	< 0.0001
Mean image score	0.36 (0.18)	1.95	0.054 (0.043[a])
DISRUPTED			
Intercept[b]	0.084 (0.85)	0.09	0.92
Mean image score[c]	−0.62 (0.27)	−2.28	0.026

[a] P_{dir} *confirming previous findings (Wedekind and Braithwaite, 2002).*

[b] *Intercept difference to control treatment.*

[c] *Slope difference to control treatment.*

punishment reputation. In their experiment, the punishment reputation turned out to be mainly built up by punishing defectors. Antisocial punishment (punishment of those who helped) was rare and declined over the course of the experiment. In the long term, the increased willingness to donate to punishers compensated for the immediate costs of punishment (dos Santos et al., 2013). Therefore, a necessary condition for the evolution of reputation-based punishment was met (dos Santos et al., 2011). Here, we exactly followed the experimental protocol of dos Santos et al. (2013) but disturbed the players (and probably constrained their working memory) by saying few superfluous words between two moves each. This additional treatment did not significantly change overall punishment rates or cooperation frequencies. Nevertheless, disturbance clearly affected punishment behavior and the subjects' discrimination between punishers and non-punishers. There was less punishment of defectors and more antisocial punishment in the disturbed environment than in the undisturbed environment. The resulting punishment reputation seemed to have lost its significance for subjects in the Donor role: we found no more effects of the punishment reputation on the likelihood of donations. Moreover, the immediate costs of punishment were not compensated during the 24 rounds in our disturbed environment, i.e., the correlation between punishment score and total income remained negative. Therefore, a key condition for the evolution of reputation-based punishment (dos Santos et al., 2011) appeared not to be met under our disturbed conditions.

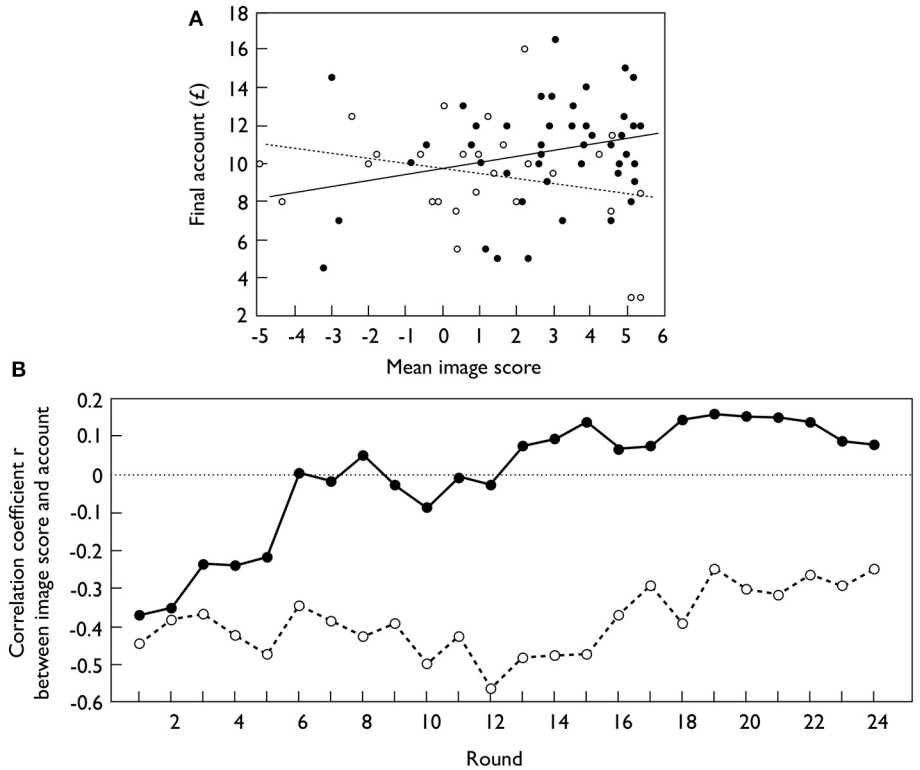

FIGURE 3 | Costs and benefits of generosity in study 1. (A) Regressions between final account and mean image score, plotted for each player in the disturbed (open symbols, hatched line) and the control condition (closed symbols, solid line). **(B)** Pearson's correlation coefficients r between image score and account at the end of each round, for the disturbed (open symbols, hatched line) and the control condition (closed symbols, solid line). See **Table 1** for statistics.

Theory predicts that simple strategies based on very little information, e.g., a simple score that summarizes previous actions, are sufficient for reputation-based cooperation to evolve (Nowak and Sigmund, 1998; Roberts, 2008; dos Santos et al., 2011). When humans were put into the respective situations and in front of a screen that provided all the information necessary to play these simple strategies, they seemed to behave as predicted: reputation based on generosity or on punishment was used, and the use of reputation increased cooperation frequencies (Wedekind and Milinski, 2000; Milinski et al., 2001; Wedekind and Braithwaite, 2002; dos Santos et al., 2013). However, such experiments cannot reveal the strategies humans use. It is still possible that humans use more sophisticated strategies than those predicted from the available models (e.g., Stevens et al., 2011), and that different strategies lead to similar outcomes under given experimental conditions. In the undisturbed conditions of our first study, for example, subjects based their decisions not only on their partners' reputation but also their own, suggesting that they managed their own reputation strategically (i.e., maintaining a positive reputation and not investing unnecessarily when the reputation is already high). Humans may also value the reputational effects of donation, defection, and punishment relative to the respective context (Barclay, 2006; Ule et al., 2009). Not donating to a Receiver with a low reputation for generosity could, for example, be valued differently to not donating to a Receiver with a high reputation for generosity (Ule et al., 2009; but see Milinski et al., 2001). The image score we provided on screen would not be differently affected by these situations, as in the original models that were tested. Moreover, it is possible that humans value more recent decisions differently to earlier ones, or that they search for behavioral patterns (e.g., analogously to the algorithm suggested by Hauert and Stenull, 2002), or that they value a given image score not against zero but against a group mean that changes over time (as for example suggested by Wedekind and Milinski, 2000 and Molleman et al., 2013). The information we provided on screen would not be sufficient to play such sophisticated strategies. The participants in the experiments would have to use more memory-demanding strategies than those proposed in the original models.

Experimentally reducing information on the others' actions and payoffs, or adding randomness to the link between an action and its payoff consequences, have been shown to decrease cooperation (Duffy and Feltovich, 1999; Bereby-Meyer and Roth, 2006). We did not directly manipulate the amount of information that was available, but by adding a disturbance and thereby probably constraining the participants' working memory, our treatment may have had a similar effect in the sense that some of the available information was discarded. The information that was discarded seems not to have been fully captured by the simple scores displayed on screen. This suggests that humans do not

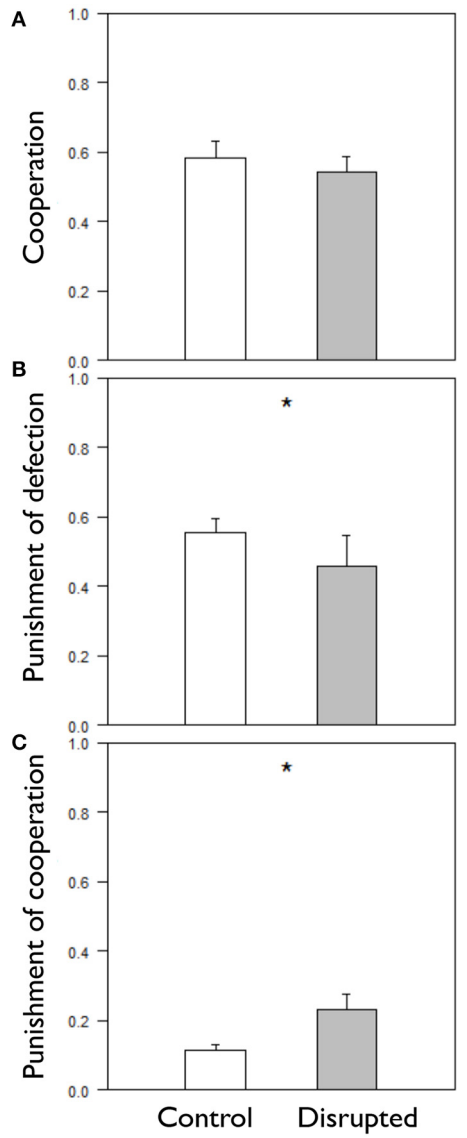

FIGURE 4 | Cooperation and punishment in the helping game with punishment (study 2). The figure shows the mean (± SE) per treatment of **(A)** cooperation, **(B)** punishment of defection, and **(C)** punishment of helping ("antisocial punishment"; $*p < 0.05$).

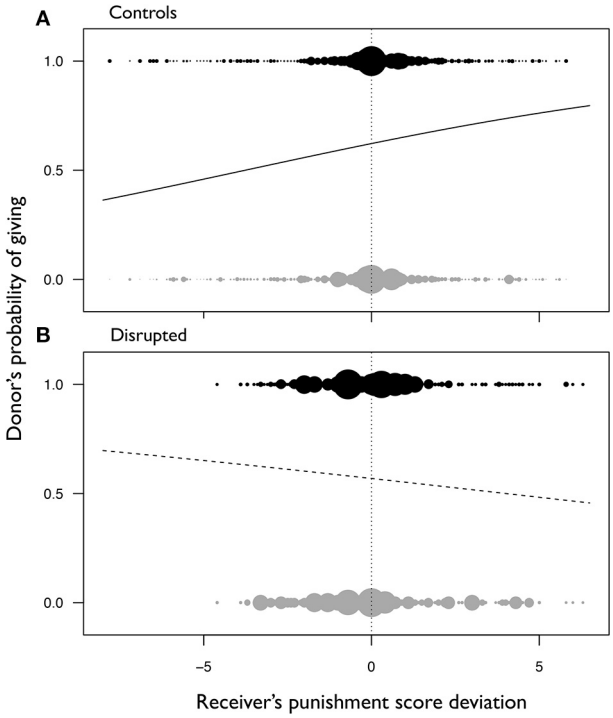

FIGURE 5 | Effects of punishment reputation on donations in (A) the control and (B) the disturbed condition in study 2. Donations (black symbols) and refused donations (gray symbols) are plotted against the Receiver's punishment reputation (deviation from group means, rounded to the nearest 0.1 so that the size of the symbols reflect the number of observations). The curves are based on GLMM estimates on the Donors' probability of giving, the dotted vertical lines indicates zero deviation from the group mean. See text for statistics.

play the simple strategies that were shown to be sufficient for the evolution of reputation-based cooperation, i.e., their strategies integrate more information than just their partners' image or punishment score (*sensu* Nowak and Sigmund, 1998; dos Santos et al., 2011). Although simple reputation games can be cooperatively solved with simple strategies in theory, real-life situations may have selected for cognitively more demanding strategies. This goes in line with the assumption that reputation use might be too cognitively demanding for most animals other than humans (Stevens et al., 2005). Reputation use by cleaner fish is a notable exception (Bshary and Grutter, 2006).

Antisocial punishment has frequently been observed in economic experiments (e.g., Herrmann et al., 2008; Nikiforakis, 2008), and theoretical studies on the evolution of antisocial punishment have demonstrated that such behavior can be expected under some conditions (Jensen, 2010; Rand et al., 2010; Rand and Nowak, 2011; Dreber and Rand, 2012). However, the finding that cooperation and punishment are altered after adding a slight disturbance to the experimental procedures implies that quantitative conclusions drawn from results of economic experiments can easily be misleading, as they would be highly dependent on how the experiments are framed and on how they are performed (Hagen and Hammerstein, 2006). Variation in the use of reputation or in punishment behavior could potentially be linked to slight variation in experimental set-ups, including the exact wording of the game instruction and possibly the appearance of the experimenter, the time of the day, or the present context of the participants (e.g., if they are students, what courses are they taking, which semester it is, whether they are close to an exam period or not). Such potentially confounding effects are very difficult to control in, for example, cross-cultural studies (e.g., Henrich et al., 2006; Herrmann et al., 2008).

We conclude that reputation-based interactions are very sensitive to disturbance. Exposure to a few superfluous words can be sufficient to significantly change behavioral strategies in reputation games. Subjects do not seem to use the simple

algorithm that was found to be sufficient for the evolution of reputation-based cooperation and punishment (Nowak and Sigmund, 1998; Leimar and Hammerstein, 2001; dos Santos et al., 2011). Quantitative differences between experimental games or groups of participants can therefore be misleading.

AUTHOR CONTRIBUTIONS

Victoria A. Braithwaite and Claus Wedekind designed the 1st study, developed the methodology, and collected the data. Miguel dos Santos and Claus Wedekind designed the 2nd study and developed the methodology. Miguel dos Santos collected the data of the 2nd study. Miguel dos Santos and Claus Wedekind analyzed all data and wrote the manuscript that was then critically revised by Victoria A. Braithwaite.

ACKNOWLEDGMENTS

We thank the students for participation, and Redouan Bshary, Laurent Lehmann, and three reviewers for useful comments on earlier versions of the manuscript. The Swiss National Science Foundation (grants to Claus Wedekind) and the John D. and Catherine T. MacArthur Foundation (Research Network on Norms and Preferences) provided financial support.

REFERENCES

Alexander, R. D. (1987). *The Biology of Moral Systems.* New York, NY: Aldine de Gruyter.

Axelrod, R., and Hamilton, W. D. (1981). The evolution of cooperation. *Science* 211, 1390–1396. doi: 10.1126/science.7466396

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evol. Hum. Behav.* 27, 325–344. doi: 10.1016/j.evolhumbehav.2006.01.003

Barclay, P. (2013). Strategies for cooperation in biological markets, especially for humans. *Evol. Hum. Behav.* 34, 164–175. doi: 10.1016/j.evolhumbehav.2013.02.002

Bates, D. M., and Sarkar, D. (2007). *lme4: Linear Mixed-Effects Models using S4 Classes*. R package version 0.99875-9.

Becker, J. T., and Morris, R. G. (1999). Working memory(s). *Brain Cogn.* 41, 1–8. doi: 10.1006/brcg.1998.1092

Bell, R., Giang, T., Mund, I., and Buchner, A. (2013). Memory for reputational trait information: is social-emotional information processing less flexible in old age? *Psychol. Aging* 28, 984–995. doi: 10.1037/a0034266

Belleville, S., Rouleau, N., Van der Linden, M., and Collette, F. (2003). Effect of manipulation and irrelevant noise on working memory capacity of patients with Alzheimer's dementia. *Neuropsychology* 17, 69–81. doi: 10.1037/0894-4105.17.1.69

Bereby-Meyer, Y., and Roth, A. E. (2006). The speed of learning in noisy games: partial reinforcement and the sustainability of cooperation. *Am. Econ. Rev.* 96, 1029–1042. doi: 10.1257/000282806779468562

Brandt, H., Hauert, C., and Sigmund, K. (2003). Punishment and reputation in spatial public goods games. *Proc. R. Soc. B Biol. Sci.* 270, 1099–1104. doi: 10.1098/rspb.2003.2336

Brosnan, S. F., Salwiczek, L., and Bshary, R. (2010). The interplay of cognition and cooperation. *Philos. Trans. R. Soc. B Biol. Sci.* 365, 2699–2710. doi: 10.1098/rstb.2010.0154

Bshary, R., and Bronstein, J. L. (2011). A general scheme to predict partner control mechanisms in pairwise cooperative interactions between unrelated individuals. *Ethology* 117, 271–283. doi: 10.1111/j.1439-0310.2011.01882.x

Bshary, R., and Grutter, A. S. (2006). Image scoring and cooperation in a cleaner fish mutualism. *Nature* 441, 975–978. doi: 10.1038/nature04755

dos Santos, M., Rankin, D. J., and Wedekind, C. (2011). The evolution of punishment through reputation. *Proc. R. Soc. B Biol. Sci.* 278, 371–377. doi: 10.1098/rspb.2010.1275

dos Santos, M., Rankin, D. J., and Wedekind, C. (2013). Human cooperation based on punishment reputation. *Evolution* 67, 2446–2450. doi: 10.1111/evo.12108

Dreber, A., and Rand, D. G. (2012). Retaliation and antisocial punishment are overlooked in many theoretical models as well as behavioral experiments. *Behav. Brain Sci.* 35:24. doi: 10.1017/S0140525X11001221

Duffy, J., and Feltovich, N. (1999). Does observation of others affect learning in strategic environments? *Exp. Study Int. J. Game Theory* 28, 131–152. doi: 10.1007/s001820050102

Earley, R. L. (2010). Social eavesdropping and the evolution of conditional cooperation and cheating strategies. *Philos. Trans. R. Soc. B Biol. Sci.* 365, 2675–2686. doi: 10.1098/rstb.2010.0147

Fu, F., Hauert, C., Nowak, M. A., and Wang, L. (2008). Reputation-based partner choice promotes cooperation in social networks. *Phys. Rev. E* 78:026117. doi: 10.1103/PhysRevE.78.026117

Gardner, A., and West, S. A. (2004). Cooperation and punishment, especially in humans. *Am. Nat.* 164, 753–764. doi: 10.1086/425623

Guala, F. (2012). Reciprocity: weak or strong? What punishment experiments do (and do not) demonstrate. *Behav. Brain Sci.* 35, 1–15. doi: 10.1017/S0140525X11000069

Hagen, E. H., and Hammerstein, P. (2006). Game theory and human evolution: a critique of some recent interpretations of experimental games. *Theor. Popul. Biol.* 69, 339–348. doi: 10.1016/j.tpb.2005.09.005

Hauert, C., and Stenull, O. (2002). Simple adaptive strategy wins the prisoner's dilemma. *J. Theor. Biol.* 218, 261–272. doi: 10.1006/yjtbi.307

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., et al. (2006). Costly punishment across human societies. *Science* 312, 1767–1770. doi: 10.1126/science.1127333

Herrmann, B., Thoni, C., and Gächter, S. (2008). Antisocial punishment across societies. *Science* 319, 1362–1367. doi: 10.1126/science.1153808

Hilbe, C., and Sigmund, K. (2010). Incentives and opportunism: from the carrot to the stick. *Proc. R. Soc. B Biol. Sci.* 277, 2427–2433. doi: 10.1098/rspb.2010.0065

Horvath, G., Kovarik, J., and Mengel, F. (2012). Limited memory can be beneficial for the evolution of cooperation. *J. Theor. Biol.* 300, 193–205. doi: 10.1016/j.jtbi.2012.01.034

Jensen, K. (2010). Punishment and spite, the dark side of cooperation. *Philos. Trans. R. Soci. B Biol. Sci.* 365, 2635–2650. doi: 10.1098/rstb.2010.0146

Larose, K., and Dubois, F. (2011). Constraints on the evolution of reciprocity: an experimental test with zebra finches. *Ethology* 117, 115–123. doi: 10.1111/j.1439-0310.2010.01850.x

Leimar, O., and Hammerstein, P. (2001). Evolution of cooperation through indirect reciprocity. *Proc. R. Soc. B Biol. Sci.* 268, 745–753. doi: 10.1098/rspb.2000.1573

Milinski, M., Semmann, D., Bakker, T. C. M., and Krambeck, H. J. (2001). Cooperation through indirect reciprocity: image scoring or standing strategy? *Proc. R. Soc. B Biol. Sci.* 268, 2495–2501. doi: 10.1098/rspb.2001.1809

Milinski, M., and Wedekind, C. (1998). Working memory constrains human cooperation in the Prisoner's Dilemma. *Proc. Natl. Acad. Sci. U.S.A.* 95, 13755–13758. doi: 10.1073/pnas.95.23.13755

Molleman, L., van den Broek, E., and Egas, M. (2013). Personal experience and reputation interact in human decisions to help reciprocally. *Proc. R. Soc. B Biol. Sci.* 280, 20123044. doi: 10.1098/rspb.2012.3044

Moreira, J., Vukov, J., Sousa, C., Santos, F. C., d'Almeida, A. F., Santos, M. D., et al. (2013). Individual memory and the emergence of cooperation. *Anim. Behav.* 85, 233–239. doi: 10.1016/j.anbehav.2012.10.030

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: can we really govern ourselves? *J. Public Econ.* 92, 91–112. doi: 10.1016/j.jpubeco.2007.04.008

Nowak, M. A., and Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature* 393, 573–577. doi: 10.1038/31225

Nowak, M. A., and Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature* 437, 1291–1298. doi: 10.1038/nature04131

Raihani, N. J., Thornton, A., and Bshary, R. (2012). Punishment and cooperation in nature. *Trends Ecol. Evol.* 27, 288–295. doi: 10.1016/j.tree.2011.12.004

Rand, D. G., Armao Iv, J. J., Nakamaru, M., and Ohtsuki, H. (2010). Anti-social punishment can prevent the co-evolution of punishment and cooperation. *J. Theor. Biol.* 265, 624–632. doi: 10.1016/j.jtbi.2010.06.010

Rand, D. G., and Nowak, M. A. (2011). The evolution of antisocial punishment in optional public goods games. *Nat. Commun.* 2:434. doi: 10.1038/ncomms1442

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rice, W. R., and Gaines, S. D. (1994). Heads I win, tails you lose: testing directional alternative hypotheses in ecological and evolutionary research. *Trends Ecol. Evol.* 9, 235–237. doi: 10.1016/0169-5347(94)90258-5

Roberts, G. (2008). Evolution of direct and indirect reciprocity. *Proc. R. Soc. B Biol. Sci.* 275, 173–179. doi: 10.1098/rspb.2007.1134

Rouleau, N., and Belleville, S. (1996). Irrelevant speech effect in aging: an assessment of inhibitory processes in working memory. *J. Gerontol. Psychol. Sci.* 51, 356–363. doi: 10.1037/0894-4105.17.1.69

Sasaki, T., and Uchida, S. (2013). The evolution of cooperation by social exclusion. *Proc. R. Soci. B Biol. Sci.* 280:20122498. doi: 10.1098/rspb.2012.2498

Sigmund, K. (2007). Punish or perish? Retaliation and collaboration among humans. *Trends Ecol. Evol.* 22, 593–600. doi: 10.1016/j.tree.2007.06.12

Sigmund, K. (2010). *The Calculus of Selfishness.* Princeton, NJ: Princeton University Press.

Sigmund, K. (2012). Moral assessment in indirect reciprocity. *J. Theor. Biol.* 299, 25–30. doi: 10.1016/j.jtbi.2011.03.024

Sommerfeld, R. D., Krambeck, H. J., Semmann, D., and Milinski, M. (2007). Gossip as an alternative for direct observation in games of indirect reciprocity. *Proc. Nat. Acad. Sci. U.S.A.* 104, 17435–17440. doi: 10.1073/pnas.0704598104

Stevens, J. R., Cushman, F. A., and Hauser, M. D. (2005). Evolving the psychological mechanisms for cooperation. *Annu. Rev. Ecol. Evol. Syst.* 36, 499–518. doi: 10.1146/annurev.ecolsys.36.113004.083814

Stevens, J. R., and Hauser, M. D. (2004). Why be nice? Psychological constraints on the evolution of cooperation. *Trends Cogn. Sci.* 8, 60–65. doi: 10.1016/j.tics.2003.12.003

Stevens, J. R., Volstorf, J., Schooler, L. J., and Rieskamp, J. (2011). Forgetting constrains the emergence of cooperative decision strategies. *Front. Psychol.* 2:235 doi: 10.3389/fpsyg.2010.00235

Sylwester, K., and Roberts, G. (2010). Cooperators benefit through reputation-based partner choice in economic games. *Biol. Lett.* 6, 659–662. doi: 10.1098/rsbl.2010.0209

Trivers, R. L. (1971). Evolution of reciprocal altruism. *Q. Rev. Biol.* 46:35. doi: 10.1086/406755

Ule, A., Schram, A., Riedl, A., and Cason, T. N. (2009). Indirect punishment and generosity toward strangers. *Science* 326, 1701–1704. doi: 10.1126/science.1178883

Volstorf, J., Rieskamp, J., and Stevens, J. R. (2011). The good, the bad, and the rare: memory for partners in social interactions. *PLoS ONE* 6:e18945. doi: 10.1371/journal.pone.0018945

Wedekind, C., and Braithwaite, V. A. (2002). The long-term benefits of human generosity in indirect reciprocity. *Curr. Biol.* 12, 1012–1015. doi: 10.1016/S0960-9822(02)00890-4

Wedekind, C., and Milinski, M. (2000). Cooperation through image scoring in humans. *Science* 288, 850–852. doi: 10.1126/science.288.5467.850

Yoeli, E., Hoffman, M., Rand, D. G., and Nowak, M. A. (2013). Powering up with indirect reciprocity in a large-scale field experiment. *Proc. Nat. Acad. Sci. U.S.A.* 110, 10424–10429. doi: 10.1073/pnas.1301210110

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 May 2014; accepted: 11 July 2014; published online: 01 August 2014.
Citation: dos Santos M, Braithwaite VA and Wedekind C (2014) Exposure to superfluous information reduces cooperation and increases antisocial punishment in reputation-based interactions. Front. Ecol. Evol. ***2****:41. doi: 10.3389/fevo.2014.00041*
This article was submitted to Behavioral and Evolutionary Ecology, a section of the journal Frontiers in Ecology and Evolution.