nature COMMUNICATIONS

ARTICLE

Received 15 Jul 2014 | Accepted 18 Sep 2014 | Published 28 Oct 2014

DOI: 10.1038/ncomms6306

Fairness violations elicit greater punishment on behalf of another than for oneself

Oriel FeldmanHall[1], Peter Sokol-Hessner[1], Jay J. Van Bavel[1] & Elizabeth A. Phelps[1,2,3]

Classic psychology and economic studies argue that punishment is the standard response to violations of fairness norms. Typically, individuals are presented with the option to punish the transgressor or not. However, such a narrow choice set may fail to capture stronger alternative preferences for restoring justice. Here we show, in contrast to the majority of findings on social punishment, that other forms of justice restoration (for example, compensation to the victim) are strongly preferred to punitive measures. Furthermore, these alternative preferences for restoring justice depend on the perspective of the deciding agent. When people are the recipient of an unfair offer, they prefer to compensate themselves without seeking retribution, even when punishment is free. Yet when people observe a fairness violation targeted at another, they change their decision to the most punitive option. Together these findings indicate that humans prefer alternative forms of justice restoration to punishment alone.

[1] Department of Psychology, New York University, New York, New York 10003, USA. [2] Center for Neural Science, New York University, New York, New York 10003, USA. [3] Nathan Kline Institute, Orangeburg, New York 10962, USA. Correspondence and requests for materials should be addressed to E.A.P. (email: liz.phelps@nyu.edu).

Social norms, such as fairness concerns, provide prescribed standards for behaviour that promote social efficiency and cooperation[1–3]. How humans resolve fairness transgressions has been extensively studied in the context of simple, constrained interactions[4]. Traditionally, people are presented with two options—engage in punitive behaviour, or do nothing. In this context, people typically respond to fairness violations with punishment[5,6]. However, such a narrow range of options may fail to capture alternative, preferred strategies for restoring justice that are frequently observed in everyday life. Here, we test alternative preferences for justice restoration by broadening the decision-making space to include compensatory measures in addition to punishment. Since impartiality is a core principle of many legal systems and is believed to influence judicial decision-making, we further test whether these preferences are differentially deployed depending on the perspective of the deciding agent. That is, do unaffected third parties sanction fairness violations differently than personally affected second parties?

Demonstrations of how intensely humans endorse punishment as a means of ensuring fair and equitable outcomes[2] suggests that punishment is the standard response to violations of justice. Hundreds of studies using the Ultimatum Game illustrate that people are willing to incur personal monetary costs to punish fairness violations. In the Ultimatum Game, two players must agree on how to split a sum of money. First, the proposer makes an offer of how to divide the money. The responder can then either accept the offer, in which case the money is split as proposed, or reject the offer, in which case neither player receives any money[7]. It is well established that responders will forgo even large monetary benefits by rejecting the offer to punish the proposer for offering an unfair split[8,9]. In fact, extremely unfair offers are rejected around 70% of the time[10].

In the real world, however, punishment is rarely the only option for restoring justice. There is a broad range of alternative responses, reflecting the idea that both the transgressor and the victim can be differentially valued depending on one's social preferences and conceptual sense of justice. For instance, some people may prefer to compensate the victim[11], or punish the transgressor such that the penalty is proportionate to the harm committed[12], preferences that may prove to have powerful roles in motivating the restoration of justice. Although existence of alternative forms of justice restoration date back as far as four millennia ago[13], no research that we are aware of has examined these alternatives alongside the prototypical punitive options.

The question of justice restoration is important because most legal systems are largely based on the principle that social order depends on punishment. For much of modern civilization, formal systems—such as judges and juries[14,15]—have been structured to mete out justice. The underlying assumption is that people make judgments differently depending on whether a fairness violation is directed towards another individual or aimed at oneself. Given the distinct asymmetries between the way people perceive themselves versus their peers[16], it is thought that unaffected and putatively dispassionate third parties sanction transgressors in a less egocentric and more deliberate manner than victims[17]. Indeed, theorists suggest that people experience psychologically close events (for example, those experienced personally) in a detailed, concrete manner, whereas socially distant objects are construed in terms of high-level, abstract characteristics and principles[18,19]. Psychological distance from a transgression may therefore bias how people evaluate fairness violations and influence their subsequent preferences for restoring justice. Accordingly, we theorized that individuals would endorse different routes to justice restoration depending on whether they are the direct recipient of a fairness violation compared with when they merely observe it.

To examine alternative motivations for restoring justice and test whether individuals navigate fairness violations differently for both self and another, we developed a novel economic game that broadens the available choice space to include a range of punitive and compensatory options for restoring justice that are not present in classic experimental games. To model alternative options for justice restoration frequently observed in the real world, we not only presented participants with the opportunity to accept or reject the proposed split (as in the Ultimatum Game), but also other novel options that reflect a range of other-regarding preferences.

In our task, Player A has the first move and can propose a division of a $10 pie with Player B (Player A: $10 − x, Player B: x, Fig. 1a). Player B can then reapportion the money by choosing from the following five options; (1) accept: agreeing to the proposed split ($10 − x, x)[7]; (2) punish: reducing Player A's payout to the original amount offered to Player B (x, x)[20]; (3) equity: equally splitting the pie so that both players receive half of the initial endowment ($5, $5)[4]; (4) compensate: increasing Player B's own payout to equal Player A's payout, thus enlarging the pie to maximizing both players' monetary outcomes ($10 − x, $10 − x)[21]; and finally, (5) reverse: reversing the proposed split—a 'just deserts' motive where the perpetrator deserves punishment

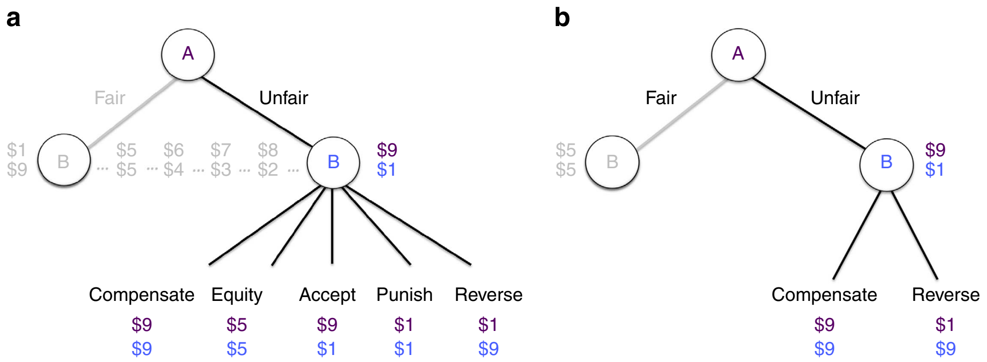


Figure 1 | Game structure. (**a**) The sequential game. Player A can make any offer to Player B. Here we illustrate all the options that Player B has to reapportion the money after being offered a split of $9/$1. On each round, however, Player B is presented with a forced choice between two options (for example, compensate versus equity, compensate versus accept, compensate versus punish, and so on) for a total of 10 pairwise comparisons. Options were randomly paired and presented across the experiment. We focused our analysis on unfair offers, splits of $6/$4 through $9/$1. (**b**) An example of a round where Player A offers Player B $1. In this case Player B is then presented with the option to either increase their own payout without decreasing Player A's payout (compensate), or reverse the payouts such that Player A receives $1 and Player B receives $9 (reverse).

proportionate to the wrong committed[12]—so that Player A is punished and Player B is compensated (x, \$10 − x)[22,23]. See Supplementary Discussion for in-depth explanations of each option. As in many classic experimental economics games that explore trade-offs between discrete choice pairs[7,24], participants were presented with only two options on any given trial, such that each option (that is, ‘compensate’, ‘equity’, ‘accept’, ‘punish’, ‘reverse’) was randomly paired with one alternative option per trial, resulting in every combination pair, for a total of 10 unique combination pairs (Fig. 1b). When making their offers, Player A was not aware which two options would be available to Player B on a given trial.

We find that although decades of research demonstrate that individuals consistently retaliate against those who behave unfairly, when alternative options for dealing with fairness violations are made available, these assumedly robust preferences to punish another are not actually preferred when offered alongside other, non-punitive options. However, when tasked with making the same decision on behalf of someone else who has experienced a fairness violation, individuals modify their responses and apply the harshest form of punishment to the transgressor. Together these results challenge our current understanding of social preferences and the emphasis placed on punitive behaviour.

Results

Preferences for justice restoration extend beyond punishment. Figure 2a shows choice behaviour ($N = 112$; 42 males, mean age 20.8 ± 2.11) for moderately unfair offers $\binom{6}{4}$ and highly unfair offers $\binom{9}{1}$ in Experiment 1. We compute endorsement rates by the frequency an option is selected, such that each option’s endorsement rate is out of 100% (number of times an option is selected/number of times the option is presented during the experiment). That is, we calculate the number of times ‘accept’ is chosen when paired with every possible alternative option, and did the same for ‘punish’, ‘compensate’, ‘equity’ and ‘reverse’. Strikingly, across all offer types, participants least chose the options ‘accept’ and ‘punish’ (10% and 16% endorsement rate, respectively; Supplementary Table 1)—the two options most similar to those in the traditional Ultimatum Game. Instead, participants most preferred the option ‘compensate’, choosing to increase their own payout and apply no punishment to Player A (92% endorsement rate; Supplementary Table 1). This preference remained robust even when participants were offered a highly unfair split of $\binom{9}{1}$ (Fig. 2a).

Since the choice pair ‘compensate’ versus ‘reverse’ controls for Player B’s monetary benefit—that is, after receiving a highly unfair spilt of $\binom{9}{1}$, choosing compensate $\binom{9}{9}$ or reverse $\binom{1}{9}$ results in the exact same monetary payout to Player B (\$9)—we can use this choice pair to directly test other-regarding preferences while controlling for Player B’s fiscal efficiency. Results reveal that when responding to unfair offers, participants prefer to compensate rather than reverse, even though punishment is free (Pearson’s $\chi^2 = 9$, 1 df, $P = 0.003$, $\varphi = 0.15$, Fig. 2b). In other words, despite the available option to maximize one’s payout while simultaneously applying punishment to Player A (selecting ‘reverse’), participants preferred to maximize their payoff and not apply any punishment to Player A. Although most previous research has focused on punishment[3] as the primary method of restoring justice, these findings illustrate that when possible, people actually prefer compensation to punishment.

In a second experiment, Player Bs were presented with varying splits of a \$1 endowment from Player A, ranging from moderately unfair $\binom{0.60}{0.40}$ to highly unfair $\binom{0.90}{0.10}$, reflected through 10 cent increments. As in Experiment 1, participants ($N = 97$, Experiment 2a) did not prefer traditional Ultimatum Game options to ‘accept’ the offer or to ‘punish’ Player A for proposing an unfair split, and instead the strongest preference was to compensate (84% endorsement rate of ‘compensate’ across all offer types, Supplementary Table 2a). Again, for unfair offers, the choice pair compensate versus reverse reveals that even when punishment is free, individuals still prefer to compensate and abstain from punishing Player A (Pearson’s $\chi^2 = 7.7$, 1 df, $P = 0.005$, $\varphi = 0.14$). Together, these findings indicate that when given the option for alternative forms of justice restoration, compensation of the victim is strongly preferred to punishment of the transgressor.

Second and third party preferences for justice restoration. To test whether being directly affected by a fairness violation influences decisions to restore justice, we also examined participants’ behaviour when they acted as a non-vested third party (Player C), observing interactions between Players A and B ($N = 261$, Experiment 2b). That is, participants were asked to

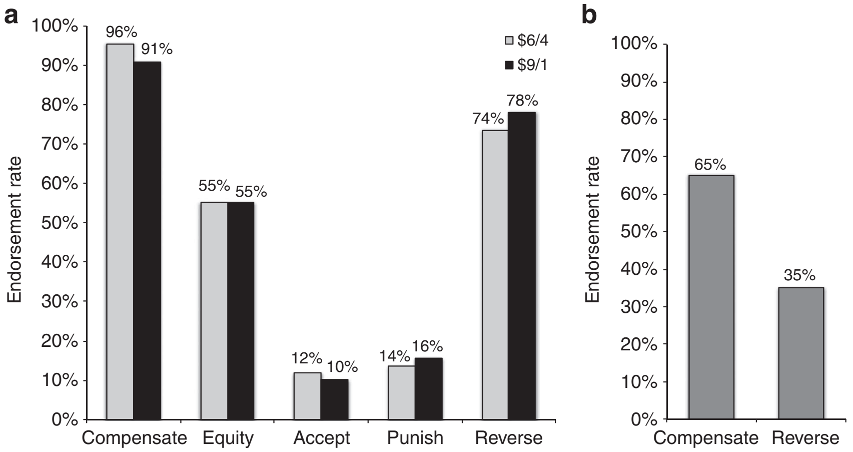


Figure 2 | Choice behaviour for restoring justice. We compute endorsement rates by the frequency an option is selected from all available trials, such that each option’s endorsement rate is out of 100%. (**a**) Results ($N = 112$) reveal that compensation is the most preferred choice, even when offered highly unfair splits. (**b**) The choice pair compensate versus reverse (game structure illustrated in Fig. 1b) equates for Player B’s fiscal efficiency, such that Player B can both compensate himself and punish Player A at no cost. Even when punishment is free, participants significantly prefer to compensate themselves and apply no punishment to Player A; Pearson’s $\chi^2 = 9$, 1 df, $P = 0.003$, $\varphi = 0.15$.

make decisions on behalf of another player such that payoffs would be paid to Players A and B and not to themselves. Unlike in the 'Self', second-party condition in which participants played the game as Player B (Experiments 1 and 2a), these 'Other', third-party decisions were non-costly and non-beneficial. Similar to decisions made in the Self condition, Player Cs (Other condition) show little preference to 'accept' the offer, or to 'punish' Player A for proposing an unfair split to Player B (Supplementary Table 2b).

Although individuals chose to compensate oneself and another at the same rate when the offer was relatively fair $\binom{0.60}{0.40}$ (McNemar's $\chi^2 = 1.2$, 1 df, $P = 0.27$), we found that when responding to unfair offers, Player Cs selected 'reverse'—the option that both compensates Player B and punishes Player A—significantly more often than Player Bs did for themselves (choice pair compensate/reverse: McNemar's $\chi^2 = 13.5$, 1 df, $P < 0.001$, $\varphi = 0.14$; Supplementary Fig. 2). In other words, although participants did not show preferences for punishing Player A when directly affected by a fairness violation (that is, as a second party), when observation of a fairness violation targeted at another (that is, as a third party), participants significantly increased their retributive responding.

Since one motive for exploring justice restoration was to investigate whether broadening the decision-making space (to include a plurality of options) affects choice behaviour, we ran four additional experiments (analysed together, see Supplementary Materials) where all five options were available on every trial. In these studies, participants were offered splits of $1 and made decisions both for themselves and on behalf of others in a within-subjects design. That is, participants made decisions both when they were personally affected by a fairness violation (as Player B; Self condition), and also on behalf of another player who was affected by a fairness violation (as Player C; Other condition).

As with our previous experiments, participants ($N = 540$) demonstrated strong preferences to 'compensate' (42% endorsement rate out of 100% across all offer types, Supplementary Fig. 3A), and did not preferentially choose to 'accept' the offer or 'punish' Player A (10% and 3% endorsement rate, respectively) when deciding for themselves. However, as the split became increasingly unfair, participants were more likely to incorporate punitive measures[17], almost doubling their endorsement of the 'reverse' option in which they simultaneously compensated themselves and punished Player A (15% endorsement of 'reverse' for relatively fair offers, compared with 30% for highly unfair offers; Cochran's Q $\chi^2 = 234$, 3 df, $P < 0.001$, Fig. 3a; analyses across all four experiments[25]). Despite this, even when offered a highly unfair split $\binom{0.90}{0.10}$, participants still preferred the least punitive and most compensatory option 'compensate' (43% endorsement rate; Cochran's Q $\chi^2 = 562.2$, 4 df, $P < 0.001$, Fig. 3a).

The participants' perspective (that is, Self versus Other condition) shifted their preferences only when the offer was highly unfair. In the Other condition, participants chose to 'reverse' the players' payouts significantly more than any other option (43% endorsement rate; Cochran's $\chi^2 = 622.2$, 4 df, $P < 0.001$; Fig. 3b, see Supplementary Fig. 3B for more details), and significantly more than they did in the Self condition (McNemar $\chi^2 = 20.2$, 1 df, $P < 0.001$, $\varphi = 0.13$, Fig. 3b). This result replicated Experiment 2, however, here participants were making decisions both as Player B and Player C (a within-subject design). Individuals who did not endorse punitive measures when deciding for themselves changed their decisions to the most retributive option after observing a fairness violation targeted at another. In contrast, there were no significant differences between choices for relatively fair offers in the Self and Other conditions (all χ^2s < 1.16, all Ps > 0.3; except for punish $\chi^2 = 4.67$, 1 df, $P = 0.03$ Fig. 3c).

Discussion

Traditionally, research has focused on punishment as the preferred response to a perceived injustice, leading to the specious assumption that people prefer to punish when righting a wrong[3,14,24,26,27]. While these studies conclude that punishment is the standard response to fairness violations, it appears that these preferences to punish may be due to a limited choice set where participants do not have the option to select from non-punitive alternatives that satisfy other preferences (for example, for equity). Here we demonstrate that when given the option to respond non-punitively to fairness violations, people derive greater utility from responding in a positive manner than they do in a punitive manner. That is, people prefer alternative forms of justice restoration, choosing compensation over punitive or retributive options. These findings fit within an emerging body of research exploring how prosocial options—like rewarding cooperation[28] so long as punishment remains a viable option[29]—can be more effective in sustaining cooperation than punishment alone.

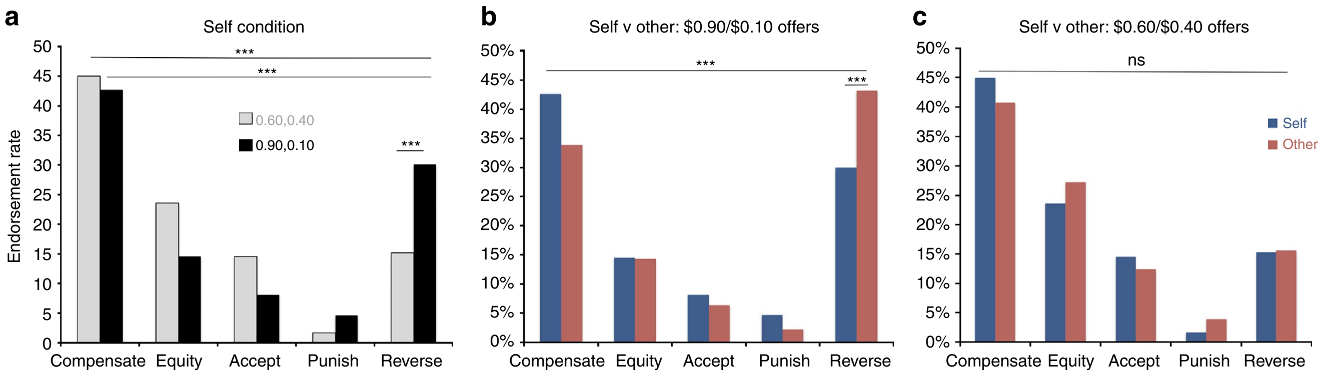


Figure 3 | Self versus Other choice behaviour. (**a**) Overall choice preferences ($n = 540$) for relatively fair offers ($0.60, $0.40) compared with highly unfair offers ($0.90, $0.10) in the Self condition: participants exhibit strong preferences for the option to compensate in both fair and unfair trials; $\chi^2 = 562.2$, 4 df, $P < 0.001$. However, preferences for retributive action become stronger when the offer is highly unfair; $\chi^2 = 234$, 3 df, $P < 0.001$. (**b**) Unfair offers ($0.90, $0.10 split) reveal that participants have significantly stronger preferences for retributive behaviour (reverse option) when making decisions for another than they do for the self; $\chi^2 = 20.2$, 1 df, $P < 0.001$, $\varphi = 0.13$. (**c**) Fair offers (0.60, 0.40 split) reveal similar choice preferences for Self and Other conditions; all χ^2s < 1.16, all Ps > 0.3; except for punish $\chi^2 = 4.67$, 1 df, $P = 0.03$. ***$P < 0.001$.

It is possible that participants chose to compensate and not punish because they prefer to maximize their own payment (rather than decrease the transgressor's payment) and because they are averse to inequality. While these are both important motivations for justice restoration, they may not necessarily be mutually exclusive. An important next question is whether people still choose to compensate even if compensation does not match Player A's payout (that is, partial compensation). Future work designed to qualitatively identify relative preferences between compensation and equality will help decipher how—and when—people trade off compensation for equality.

There are of course instances when punishment becomes a more attractive response than non-punitive options. Depending on the options punishment is juxtaposed against, deciding to punish may provide the greatest utility. For example, when offered alongside the option to accept an unfair offer, punishment (for example, equalizing both players' payoffs as well as reducing the payoff of the transgressor) is the most preferred option in our experiments and in the abundant research employing the Ultimatum Game. Combining our findings with prior research on punishment clearly demonstrates that the preference for punishment can be differentially valued depending on the landscape of options. Punishment, compensation, equity, and other alternatives to justice restoration may all provide varying degrees of utility depending on the alternative available options and the extent of the fairness violation in the first place. However, the evidence that people exhibit strong preferences to compensate when responding to fairness violations suggests that the current emphasis on punishment fails to capture other important alternatives for justice restoration.

Interestingly however, when responding to a fairness violation on behalf of another, individuals shift their preferences for restoring justice to include the most punitive and retributive measures. That individuals prefer more punitive options when deciding on behalf of another but not for oneself illustrates that context can dramatically alter the attractiveness of punishment as a measure of justice restoration. One possible explanation for the observed differences in choice behaviour between Self and Other is that deciding for another entails greater psychological distance. Increasing psychological distance—including social distance—emphasizes higher-level, abstract characteristics in the perception, experience and evaluation of situations or objects[19,30]. When deciding on behalf of another, people may be attending to schematic representations of justice—abstract ideological values such as 'justice as fairness'[31]—which emphasizes the application of known social norms to right a perceived injustice. In this case, punitive responding increases because people can easily rely on the straightforward prescriptions of punishing as a means to restore justice. On the other hand, when making decisions for oneself, events may be construed in terms of low-level, concrete and essential features, including the possibility of monetary gain. When directly experiencing a fairness violation, people may be ignoring the straightforward prescriptions of justice (to punish), instead concretely evaluating each option and its consequences. Thus, the focus is less on punishing the transgressor and more on compensating for oneself.

Here we illustrate that when presented with alternative options for restoring justice, people do not prefer to punish. We also demonstrate that people respond more punitively on behalf of others than they do for themselves. The findings that victims prefer compensation over punishment could inform how the legal system approaches the punishment of transgressors. How to restore justice is a complex question, and while this research is only an initial step, it highlights the myopia of our understanding to date, and the critical importance of considering alternative means of making what was wrong, right.

Methods

Experiment 1. Experiment 1 was run at the laboratory of the Center for Experimental Social Science (CESS) at New York University. One hundred and twelve participants participated, drawn from the general undergraduate population and recruited through e-mail solicitations. Each experimental session lasted ~1 h. All experiments were approved by New York University's Committee on Activities Involving Human Subjects and all participants completed a consent form before starting the experiment.

We utilized a pairwise comparison design that allowed us to directly contrast every choice pair (as in the Ultimatum Game, Fig. 1b). We recruited as many as 22 participants during one session, randomly assigning half of the participants to play as Player A and the other half to play as Player B for the duration of the entire experiment. All participants were paid an initial $10 show-up fee and an additional bonus depending on their choices (ranging from $1 to $9), which falls within the traditional monetary incentive structure for Ultimatum Games[32]. The instructions were read out loud so that all participants were collectively made aware of the rules. Full instructions can be found in the Supplementary Materials. On each trial, participants were randomly and anonymously paired with other participants in the room, resulting in 70 one-shot games. On every trial, all Player As were endowed with $10 and were told to make a split in whatever way he or she sees fit with Player B, so long as it is in whole dollar increments. Player B was then presented with options to reapportion the money. Altogether there were five options, however, only two of these options were presented at one time on any given trial (Fig. 1b). Participants were made aware that options to reapportion the money would be randomly paired and presented on each trial. Furthermore, participants were told that one trial would be randomly selected to be paid out and that half of the time the trial would be paid out according to Player A's split (like a dictator game), and half the time according to the decision by Player B to reapportion the money (see Supplementary Methods for more task details). Although Player A could choose to split the money in whatever way they saw fit, our aim was to understand social preferences for restoring justice, and so we restricted our analysis to unfair splits of $10, ranging from moderately unfair $\binom{6}{4}$ to highly unfair $\binom{9}{1}$.

Experiments 2–6. Participants were recruited from the United States using the online labour market Amazon Mechanical Turk (AMT)[33–36]. Participants played anonymously over the Internet and were not allowed to participate in more than one experimental session. On each trial, participants (Player B) were paid an initial participation fee of $0.50 and an additional bonus depending on their choices (ranging from $0.10 to $0.90). Across all experiments, participants were first presented with a standard digital consent form, which explained the general procedure, known risks (none), confidentiality, compensation and their rights. They could only partake in the study once they agreed to the consent form.

To ensure task comprehension, participants had to correctly complete a quiz following the instructions. Only after they correctly completed the quiz could participants begin the task. Participants were then told to place their hands on the keyboard on the following keys: S, D, F, H, J, and a timer counted down from five before the task started. On each trial, the options 'compensate', 'equity', 'accept', 'punish' and 'reverse' (labelled in analyses and here, but not presented to participants; see Supplementary Fig. 4) were displayed in a different order. After completing the task, participants were explicitly probed on their strategies when the offer was relatively fair $\binom{0.60}{0.40}$ and when the offer was highly unfair $\binom{0.90}{0.10}$, for both the Self and Other conditions. That is, participants were asked 'in your own words please describe your strategy for a scenario when Player A kept $0.60 and offered $0.40 to you'. See Supplementary Materials for a sampling of participants' strategies.

Unlike the experiments run in the laboratory, in the experiments run through AMT, we restricted offers from Player A (in reality, predetermined offers from a computer) to varying levels of unfairness, ranging from moderately unfair $\binom{0.60}{0.40}$ to highly unfair $\binom{0.90}{0.10}$, reflected through $0.10 increments. This was done primarily because we were interested in how people resolve fairness transgressions.

Differences in task structure for experiments 2–6. Experiment 2 was a pairwise comparison of each choice pair (Fig. 1b). Participants ($N=358$) played the task either as Player B (Self condition; $N=97$) or as Player C (Other condition; $N=261$), a between-subjects design. Participants were only instructed about the condition they were in, such that the instructions either explained that participants were to make decisions for themselves and Player A (Self condition), or on behalf of two other Players (Other condition). Participants were able to make an additional payout based on their choices if they completed the Self condition. For participants who completed the Other condition, they did not make an additional bonus but were paid for the time taken to complete the task.

Like in Experiment 1, on each trial, participants were presented with only two options. For example, after being offered an unfair split, Player B only observed two options (for example, compensate versus equity, compensate versus accept, compensate versus punish, compensate versus reverse, equity versus accept, equity versus punish, and so on). Thus, for every offer type ($(\binom{0.60}{0.40}$—$\binom{0.90}{0.10})$), participants saw all possible pairwise comparisons (that is, 10 pairs for each offer type, and four different offer types, resulting in 40 anonymous, one-shot games in total). Trials were randomly presented to participants.

In Experiments 3–6, participants played the task as both Player B and Player C. This within-subject design allowed us to explore each individual's choices across conditions, Self and Other. Although Experiments 3–6 were quite similar, there were small differences between the tasks which are enumerated here. In Experiment 3, Self and Other trials were presented in discrete blocks, with the Self condition always presented first and the Other condition presented second. However, to ensure that there were no order effects and that participants were not anchoring their decisions according to the decisions made in the first block (Self condition), Experiments 4–6 randomly presented the trials such that Self and Other trials were randomly interleaved across the experiment. In Experiment 3, reaction times were collected with a mouse, whereas in Experiments 4–6, reaction times were collected using the keyboard (button presses). Reaction time data were similar regardless of whether participants used a mouse or a keyboard: across all four Experiments, participants were faster to decide for another than they were for themselves (see reaction time data in Supplementary Materials). In Experiment 4, each participant was presented with a random ordering of trials. In other words, no participant saw the same order of offer types. In Experiment 5, all participants were presented with the same randomized set of trials. That is, AMT presented the same order of trials (previously determined by an algorithm to randomly interleave offer types and conditions) to all participants. Experiment 6 followed the same structure as Experiment 5, with the only difference being that blank profile pictures were added to the instructions to further delineate the roles of all the players.

References

1. Falk, A., Fehr, E. & Fischbacher, U. Testing theories of fairness—Intentions matter. *Games Econ. Behav.* **62,** 287–303 (2008).
2. Fehr, E. & Fischbacher, U. Why social preferences matter—The impact of non-selfish motives on competition, cooperation and incentives. *Econ. J.* **112,** C1–C33 (2002).
3. Fehr, E., Fischbacher, U. & Gachter, S. Strong reciprocity, human cooperation, and the enforcement of social norms. *Hum. Nature* **13,** 1–25 (2002).
4. Fehr, E. & Schmidt, K. M. A theory of fairness, competition, and cooperation. *Q. J. Econ.* **114,** 817–868 (1999).
5. Herrmann, B., Thoni, C. & Gachter, S. Antisocial punishment across societies. *Science* **319,** 1362–1367 (2008).
6. Fowler, J. H. Altruistic punishment and the origin of cooperation. *Proc. Natl Acad. Sci. USA* **102,** 7047–7049 (2005).
7. Guth, W., Schmittberger, R. & Schwarze, B. An experimental-analysis of ultimatum bargaining. *J. Econ. Behav. Organ.* **3,** 367–388 (1982).
8. Cameron, L. A. Raising the stakes in the ultimatum game: Experimental evidence from Indonesia. *Econ. Inq.* **37,** 47–59 (1999).
9. Slonim, R. & Roth, A. E. Learning in high stakes ultimatum games: an experiment in the Slovak Republic. *Econometrica* **66,** 569–596 (1998).
10. Camerer, C. *Behavioral Game Theory: Experiments in Strategic Interaction* (Russell Sage Foundation; Princeton University Press, 2003).
11. Weitekamp, E. Reparative Justice: towards a victim oriented system. *Eur. J. Criminal Policy Res.* **1,** 70–93 (1993).
12. Carlsmith, K. M., Darley, J. M. & Robinson, P. H. Why do we punish? Deterrence and just deserts as motives for punishment. *J. Pers. Soc. Psychol.* **83,** 284–299 (2002).
13. Gurney, O. R. & Kramer, S. N. *Two Fragments of Sumerian Laws* (University of Chicago Press, 1965).
14. Fehr, E. & Fischbacher, U. Third-party punishment and social norms. *Evol. Hum. Behav.* **25,** 63–87 (2004).
15. Smith, A. *The Theory of Moral Sentiments* (A. Millar, 1759).
16. Nisbett, R. E., Legant, P. & Marecek, J. Behavior as seen by actor and as seen by observer. *J. Pers. Soc. Psychol.* **27,** 154–164 (1973).
17. Fehr, E. & Fischbacher, U. Social norms and human cooperation. *Trends Cogn. Sci.* **8,** 185–190 (2004).
18. Ledgerwood, A., Trope, Y. & Liberman, N. Flexibility and consistency in evaluative responding: the function of construal level. *Adv. Exp. Soc. Psychol.* **43,** 257–295 (2010).
19. Trope, Y. & Liberman, N. Construal-level theory of psychological distance. *Psychol. Rev.* **117,** 440–463 (2010).
20. Bolton, G. E. & Zwick, R. Anonymity versus punishment in ultimatum bargaining. *Games Econ. Behav.* **10,** 95–121 (1995).
21. Lotz, S., Okimoto, T. G., Schlosser, T. & Fetchenhauer, D. Punitive versus compensatory reactions to injustice: Emotional antecedents to third-party interventions. *J. Exp. Soc. Psychol.* **47,** 477–480 (2011).
22. Pillutla, M. M. & Murnighan, J. K. Unfairness, anger, and spite: Emotional rejections of ultimatum offers. *Organ. Behav. Hum. Decis. Process* **68,** 208–224 (1996).
23. Straub, P. G. & Murnighan, J. K. An experimental investigation of ultimatum games—information, fairness, expectations, and lowest acceptable offers. *J. Econ. Behav. Organ.* **27,** 345–364 (1995).
24. Fehr, E. & Gachter, S. Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* **90,** 980–994 (2000).
25. Schimmack, U. The ironic effect of significant results on the credibility of multiple-study articles. *Psychol. Methods* **17,** 551–566 (2012).
26. Henrich, J. *et al.* 'Economic man' in cross-cultural perspective: behavioral experiments in 15 small-scale societies. *Behav. Brain Sci.* **28,** 795–815 (2005).
27. Rand, D. G. & Nowak, M. A. The evolution of antisocial punishment in optional public goods games. *Nat. Commun.* **2,** 434 (2011).
28. Rand, D. G., Dreber, A., Ellingsen, T., Fudenberg, D. & Nowak, M. A. Positive interactions promote public cooperation. *Science* **325,** 1272–1275 (2009).
29. Andreoni, J., Harbaugh, W. & Vesterlund, L. The carrot or the stick: rewards, punishments, and cooperation. *Am. Econ. Rev.* **93,** 893–902 (2003).
30. Ledgerwood, A., Trope, Y. & Chaiken, S. Flexibility now, consistency later: psychological distance and construal shape evaluative responding. *J. Pers. Soc. Psychol.* **99,** 32–51 (2010).
31. Rawls, J. *Theory of Justice* 38–52 (Harvard University, 1994).
32. Camerer, C. T. & Thaler, R. H. Anomalies: ultimatums, dictators and manners. *J. Econ. Behav. Perspect.* **9,** 209–219 (1995).
33. Mason, W. & Suri, S. Conducting behavioral research on Amazon's Mechanical Turk. *Behav. Res. Methods* **44,** 1–23 (2012).
34. Horton, J. J., Rand, D. G. & Zeckhauser, R. J. The online laboratory: conducting experiments in a real labor market. *Exp. Econ.* **14,** 399–425 (2011).
35. Paolacci, G., Chandler, J. & Ipeirotis, P. G. Running experiments on Amazon Mechanical Turk. *Judgm. Decis. Making* **5,** 411–419 (2010).
36. Buhrmester, M., Kwang, T. & Gosling, S. D. Amazon's mechanical turk: a new source of inexpensive, yet high-quality, data? *Perspect. Psychol. Sci.* **6,** 3–5 (2011).

Acknowledgements

We are grateful to Dean Mobbs and Tim Dalgleish for their early help and support with this research. This research is supported by a grant from the National Institute of Aging.

Author contributions

O.F.H. designed the experiments in consultation with E.A.P. and P.S.H. O.F.H. and P.S.H. carried out the experiments. O.F.H. ran the statistical analyses, and O.F.H., P.S.H., J.J.V.B. and E.A.P. wrote the paper.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

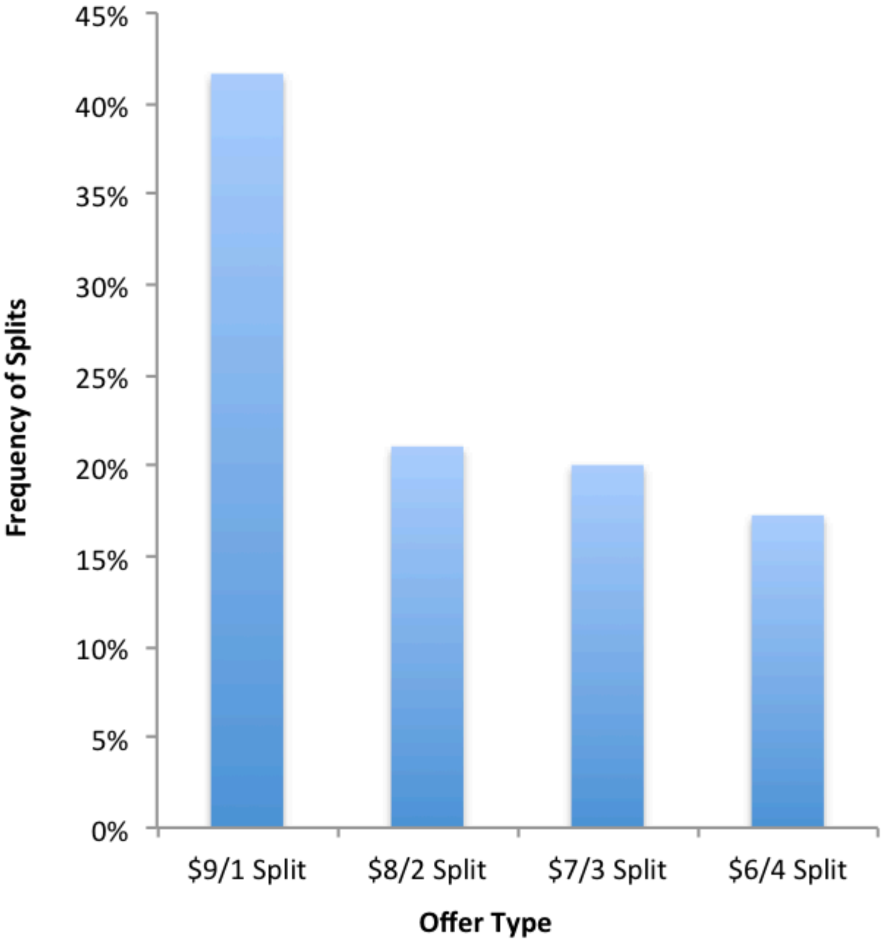
Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

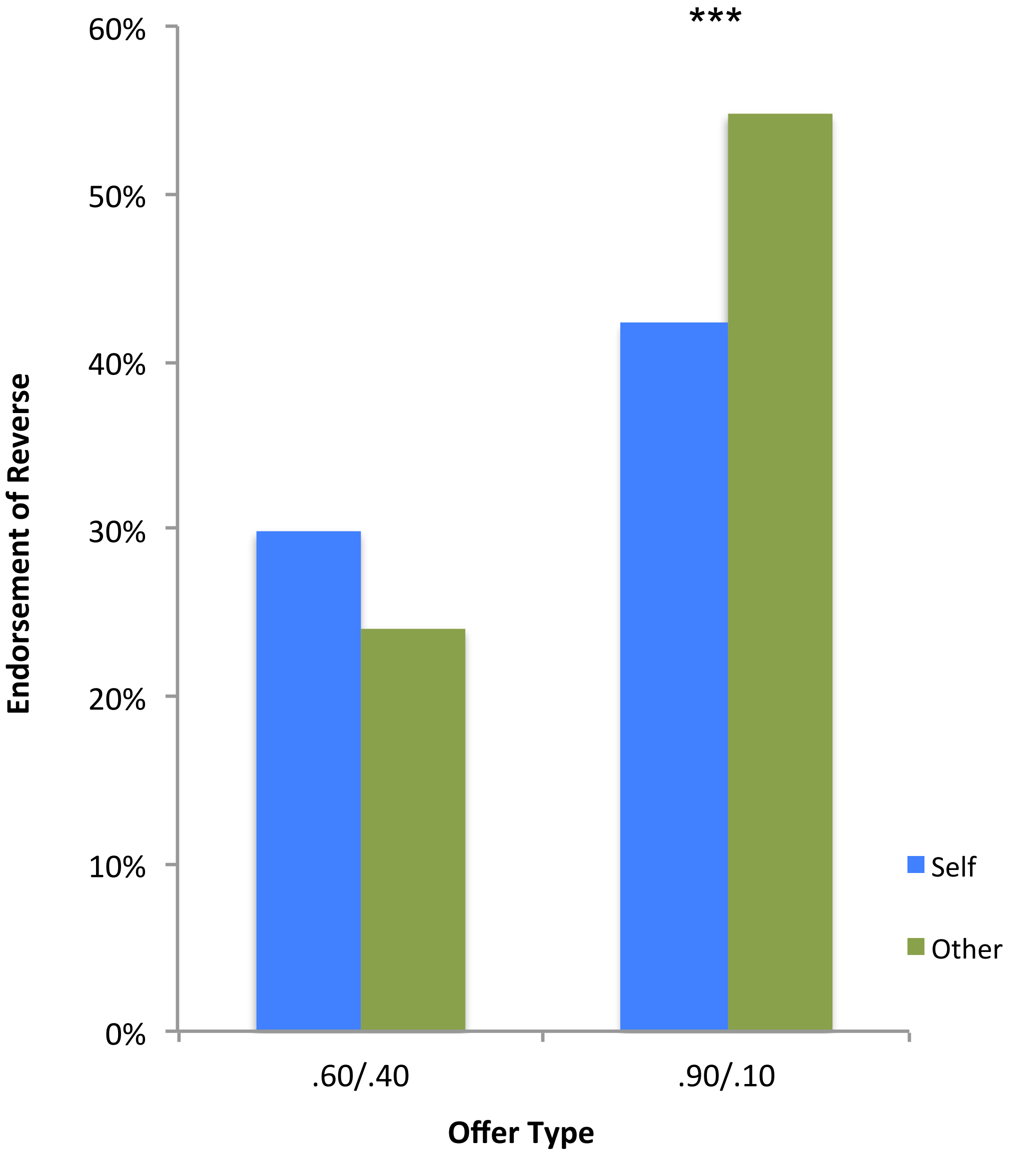
How to cite this article: FeldmanHall, O. *et al.* Fairness violations elicit greater punishment on behalf of another than for oneself. *Nat. Commun.* 5:5306 doi: 10.1038/ncomms6306 (2014).

SUPPLEMENTARY MATERIALS

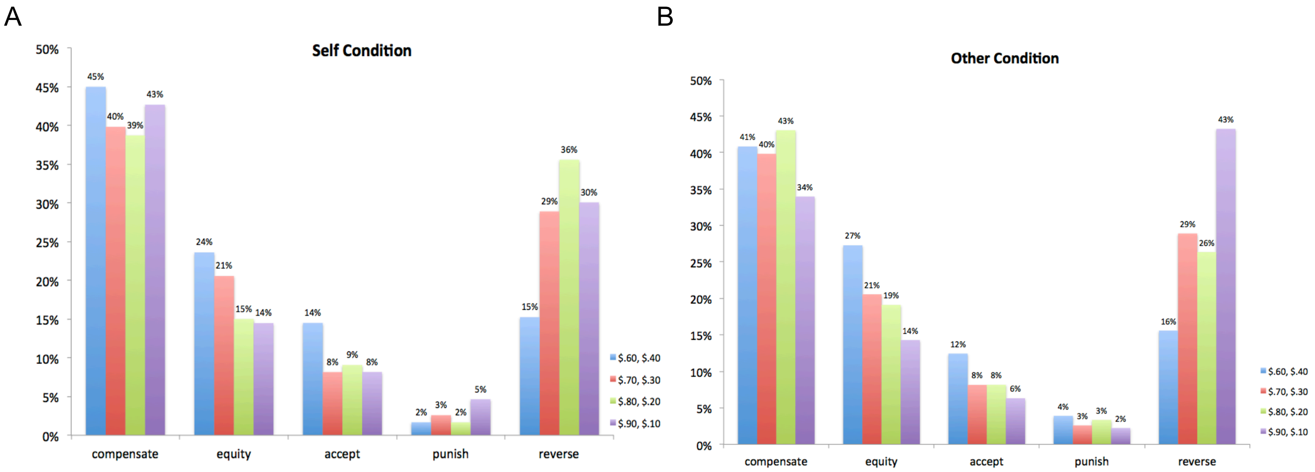
SUPPLEMENTARY FIGURES



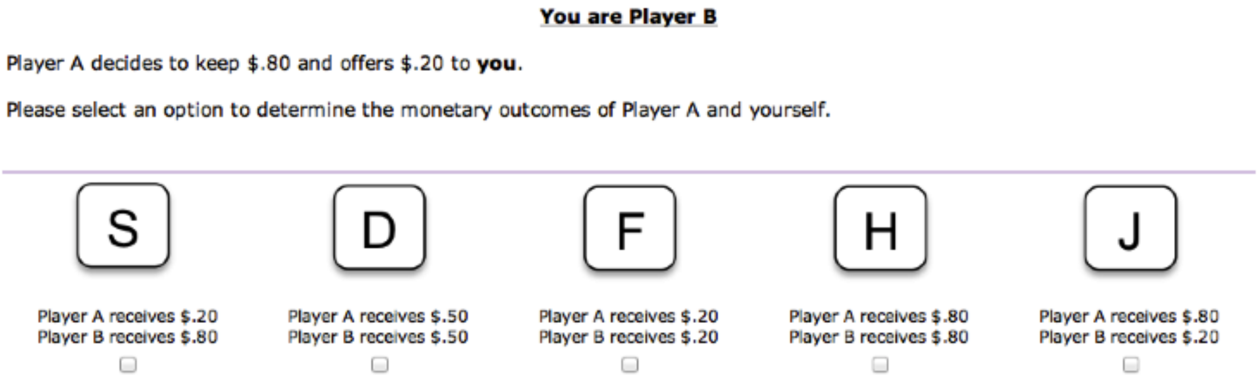
Supplementary Figure 1 | Frequency of unfair splits from Player A in Experiment 1. Player As made highly unfair offers of ($9/$1) 42% of the time.



Supplementary Figure 2 | Choice Behavior for Experiment 2. Endorsement rates of each option paired with all possible other options in the Self condition (Experiment 2a) and Other condition (Experiment 2b).



Supplementary Figure 3 | Choice Behavior for Experiments 3-6. A) Endorsement rates of each option in the Self condition, illustrating that regardless of the offer type, participants prefer to compensate and not punish. **B).** Endorsement rates of each option in the Other condition, illustrating that when the offer becomes unfair, participants significantly prefer to reverse the payouts on behalf of another.



You are Player B

Player A decides to keep $.80 and offers $.20 to **you**.

Please select an option to determine the monetary outcomes of Player A and yourself.

Player A receives $.20
Player B receives $.80
☐

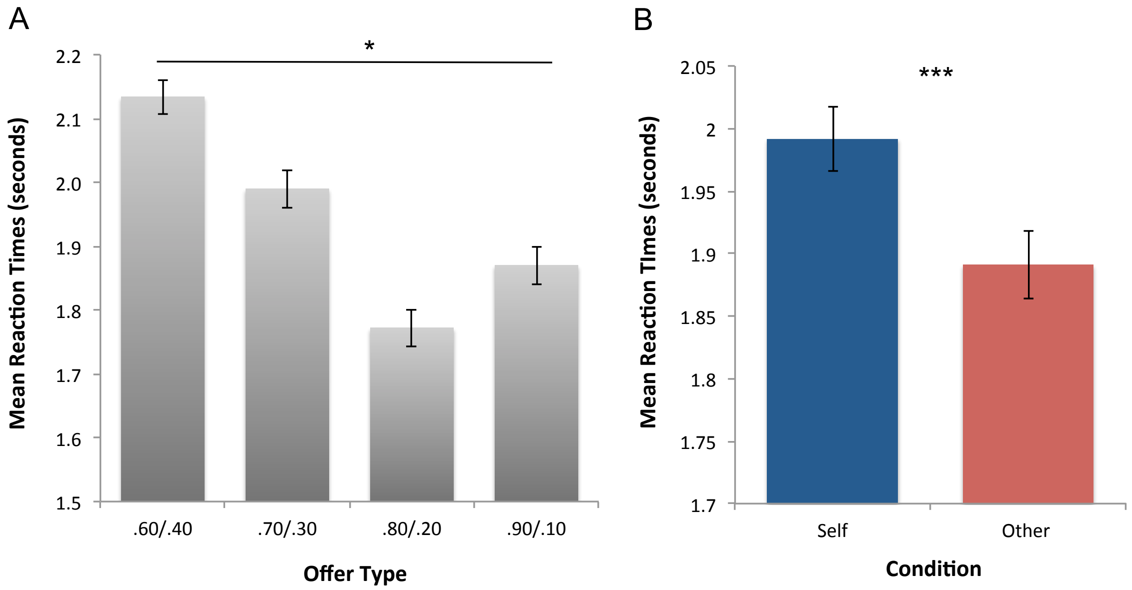
Player A receives $.50
Player B receives $.50
☐

Player A receives $.20
Player B receives $.20
☐

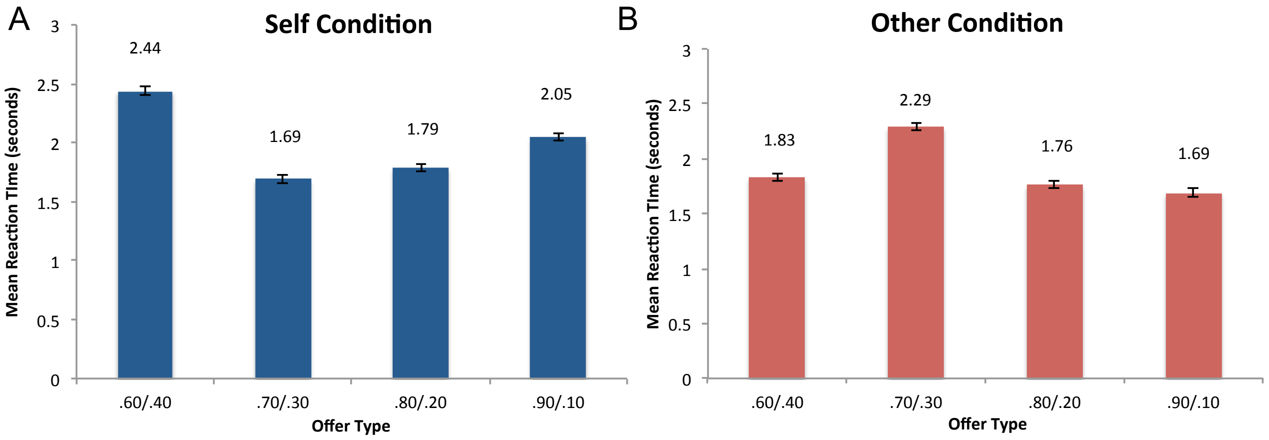
Player A receives $.80
Player B receives $.80
☐

Player A receives $.80
Player B receives $.20
☐

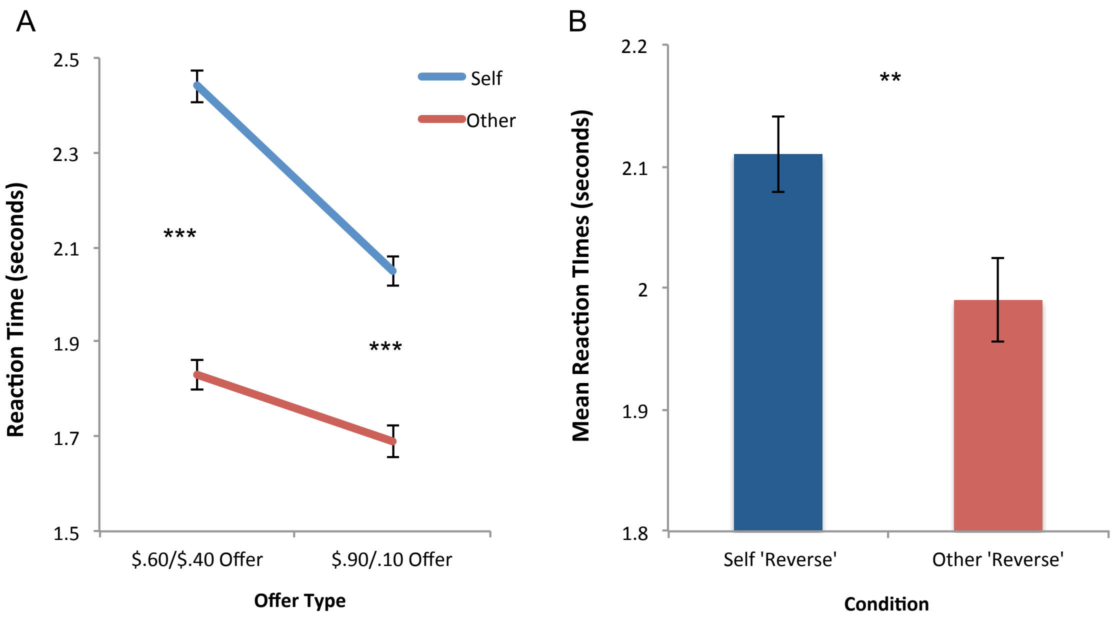
Supplementary Figure 4 | **Example Trial.** Visual of a trial in the Self condition where Player A offers an $0.80/$0.20 split to Player B. By pressing one of the five buttons, participants were able to determine the monetary outcomes for themselves and Player A.



Supplementary Figure 5 | Reaction Times by Offer Type. A) Reaction time responses (regardless of what the endorsed option) for each offer type, conditions collapsed. **B)** Reaction times for all response types in 'Self' and 'Other' conditions, all offer types collapsed. ***p<0.001, *p<0.05. Error bars represent 1 SEM.



Supplementary Figure 6 | Reaction Time by Condition. A) Reaction times in Self Condition**. B)** Reaction times in Other Condition. Error bars represent 1 SEM.



Supplementary Figure 7 | Reaction Times. A) Mean reaction times by condition and offer type ($.60, $.40 and $.90, $.10) illustrate that participants make significantly faster, more retributive decisions for another when the offer is unfair, compared to the slower, more prosocial choices made for the self. Error bars indicate one standard error of the mean. **B)** Mean reaction times for the option to 'reverse' reveal that participants are significantly slower to be retributive when deciding for themselves compared to when deciding on behalf of another. ***p<0.001 ** p<0.01

SUPPLEMENTARY TABLES

$9/1 Split	COMPENSATE	EQUITY	ACCEPT	PUNISH	REVERSE	TOTAL
compensate		99%	99%	100%	65%	91%
equity	1%		99%	100%	21%	55%
accept	1%	1%		37%	2%	10%
punish	0%	0%	63%		0%	16%
reverse	35%	79%	98%	100%		78%
						Total Trials: 1052
$8/2 Split	**compensate**	**equity**	**accept**	**punish**	**reverse**	**Total**
compensate		100%	100%	100%	53%	88%
equity	0%		97%	100%	20%	54%
accept	0%	3%		33%	0%	9%
punish	0%	0%	67%		0%	17%
reverse	47%	80%	100%	100%		82%
						Total Trials: 532
$7/3 Split	**compensate**	**equity**	**accept**	**punish**	**reverse**	**Total**
compensate		100%	100%	100%	75%	94%
equity	0%		100%	100%	4%	51%
accept	0%	0%		28%	2%	8%
punish	0%	0%	72%		0%	18%
reverse	25%	96%	98%	100%		80%
						Total Trials: 506
$6/4 Split	**compensate**	**equity**	**accept**	**punish**	**reverse**	**Total**
compensate		100%	100%	100%	82%	96%
equity	0%		97%	100%	24%	55%
accept	0%	3%		45%	0%	12%
punish	0%	0%	55%		0%	14%
reverse	18%	76%	100%	100%		74%
						Total Trials: 436

Supplementary Table 1 | Endorsement rates of each option when paired with every possible pairwise option. Endorsement of each option is designated on the left (Y axis) and paired with every possibility on the right (X axis).

A

SELF	$.60/.40	$.70/.30	$.80/.20	$.90/.10	TOTAL
Compensate	86%	84%	85%	82%	84%
Equity	66%	64%	63%	63%	64%
Accept	11%	11%	10%	9%	10%
Punish	26%	29%	28%	27%	27%
Reverse	61%	62%	64%	69%	64%

B

OTHER	$.60/.40	$.70/.30	$.80/.20	$.90/.10	TOTAL
Compensate	70%	63%	61%	59%	63%
Equity	71%	73%	73%	72%	72%
Accept	23%	15%	12%	11%	15%
Punish	43%	45%	44%	43%	44%
Reverse	42%	55%	60%	64%	55%

Supplementary Table 2 | Endorsement rates of each option paired with all possible other options. A). Self condition (Experiment 2a) and **B)**. Other condition (Experiment 2b).

SUPPLEMENTARY DISCUSSION

Motivations for Restoring Justice

According to rational choice theory[1], individuals are motivated by material self-interest, always optimizing the expected utility of options when making decisions[2]. Yet decades of work exploring how people respond to fairness violations suggest that there are strong motivational forces that drive deviations from economic self-interest[3]. Such departures from self-interest have inspired models of social preferences, such as reciprocal fairness, where players are assumed to positively value kind intentions, and to negatively value hostile intentions[3]. For example, if player A reduces B's payoff to his own benefit, a reciprocal player B will punish A, whereas if the reduction of player B's payoff was a result of a unintentional redistribution, player B will not punish A[4]. Alternatively, if a player is motivated by inequity aversion, or the dislike of unequal outcomes[3], then player B will take action to redistribute income[5].

In these classic decision-making games, motives of punishment, inequality aversion, and cooperation are pitted against a singular other motive. In an attempt to understand whether punishment and compensation are psychologically similar approaches to restoring justice, we have devised a novel economic game in which participants have multiple options for restoring justice, each of which harnesses a different motivation. Below we explain in detail the rationale behind each option.

Accept: Accepting an offer from Player A reflects a classic option in the literature [6]. When accepting an offer from Player A, Player B is typically agreeing to receive a smaller amount relative to what Player A apportions for him or herself.

Punish: Although choosing to punish in the Ultimatum Game traditionally requires participants to select the option where neither player receives any money ($0, $0)[7], we modified this option to allow for minor fiscal payout. We rationalized that punishing Player A by dropping their payout to equal the amount offered to Player B was a moderate form of punishment not resulting in a null payout for either player. In this case, Player B's payout is not altered, and instead Player A's payout is reduced to match the initial offer to Player B.

Reverse: According to the theory of retributive justice, the most appropriate response is to ensure that punishment is proportionate to the crime committed. Retributive justice is as old as recorded history, and is enshrined within legal documents and cultures around the world. These philosophies have been formalized in classic psychological theory: if the punishment fits the crime, a person is deservingly punished proportionate to the moral wrong committed. This is typically referred to as a 'just deserts' or deservingness principle[8]. In order to operationalize this in our task, we reasoned that reversing the Players' outcomes allows for the maximum punishment to be applied to Player A while also giving the maximum compensation to Player B. Moreover, reversing the Players' payouts results in Player A receiving what was initially assigned for Player B, and vice versa—a direct implementation of the 'just deserts' principle.

Compensate: While most modern societies endorse punishment as a standard practice for restoring justice, in some primitive societies, in lieu of punishing the criminal, justice could be restored by providing monetary compensation to the victim[9]. More recently, research has also demonstrated that people have strong social preferences for equitable and efficient outcomes that increases the payouts of all recipients[10]. Indeed, theories of fairness[3] predict that people may have a preference to compensate rather than punish. Given this, we operationalized 'compensation' as increasing the victim's (Player B) monetary payout without decreasing Player A's payout (the Pareto efficient option). While this option increases the total monetary pie—such that Player A and Player B can both receive more money then was initially endowed to

Player A—there are many examples in the real world where such scenarios transpire. For instance, when filing an insurance claim for stolen goods, it is unlikely that the stolen goods will be recovered and recouped by the victim. Because of this, the insurance company provides monetary compensation to cover the stolen goods. In this case, both the criminal and the victim end up with increased fiscal benefit.

Equity: This option reflects two motivations that are not mutually exclusive. First, the option to equally distribute the payouts ($5, $5) allows for a moderate amount of compensation for the victim and a moderate amount of punishment to be applied to the transgressor. This option allows participants to balance a desire to both compensate and punish. Second, in much the same way that 'compensating' distributes equal payouts to both players, the 'equity' option also controls for participants' putative aversion to inequality[3].

Experiment 1 Choice Data

We plot the data for all unfair offer types (Figure S1). Player As routinely offered highly unfair splits of $\binom{9}{1}$. Regardless of how unfair the offer from Player A is, Player Bs prefer to compensate and apply no punishment to Player A. Table S1 delineates the endorsement of each option compared to every other option for each offer type (pairwise comparisons). For example, for a $\binom{9}{1}$ split, participants chose to compensate 99% of the time when the other presented option was equity, 99% when the other option presented was accept, 100% when the other option presented was punish, and 65% of the time when other option presented was reverse.

Experiment 1 Strategies

After finishing the experiment, we asked all participants to describe in their own words their strategy used during the game. Below we include a handful of representative comments from Player A.

- *"I always selected the highest payoff for me."*
- *"I felt kinda bad doing $1 for B, so I did $2. I was hoping by not giving the absolute minimum they would show mercy to me if they to choose between lowering my pay or accepting the offer."*
- *"Max payout for myself"*
- *"I gave B as little as possible and hoped B's options were in my favor"*

Below we include a handful of representative comments from Player B.

- *"I always chose the profitable option while trying not to hurt Player A"*
- *"I picked the option that was best for both of us, unless I was going to make a significantly less amount than the other player"*
- *"I picked whichever gave me the most money while also trying to benefit role A if I could"*
- *"I was Player B, so usually I selected the option that benefited [sic] both players"*
- *"I picked the highest amount for myself. If both options were to yield the same payout for me I picked what gave (player) A the most"*

Experiments 3-6 Choice Data

Figure S2A illustrates participants' responses across all offer types when deciding for themselves. Although we found significant preferences for 'compensate' compared to every other option across all offer levels (X^2s > 11.79, 1df, Ps < 0.001 analyses across all four experiments [18]), participants' preferences also depended on what type of offer they received. As the offer became increasingly unfair, participants preferentially chose to 'reverse' the outcomes, an option that simultaneously compensates themselves and punishes Player A. When deciding for another (Other condition), participants exhibit similar behavior for most offer types ($\binom{.60}{.40}$ splits –$\binom{.80}{.20}$ splits). However, when the offer became highly unfair $\binom{.90}{.10}$, participants shifted their behavior remarkably, such that the 'reverse' option became the most preferred response (Fig S3B).

Directly comparing responses between the Self and Other condition for relatively fair offers ($.60, $.40) compared to highly unfair offers $\binom{.90}{.10}$ reveals differential behavior across the two conditions, such that participants chose the most retributive option ('reverse') significantly more when deciding for another when the offer is highly unfair (see manuscript for analysis). However, directly comparing responses between the 'Self' and 'Other' conditions for $.60, $.40 and $.70, $.30 offers, illustrate remarkably similar results between the two conditions (X^2=4.0, 4df, p=.40). This suggests then when presented with relatively fair offers, participants appear to process these offers in a relatively similar fashion for both themselves and others.

Experiments 3-6 Reaction Time Data

To help understand the cognitive mechanisms underlying choice behavior to restore justice, we examined the speed (reaction times) with which participants made their choices in Experiments 3-6. Because analyzing reaction time data in a between group design has many pitfalls, including difficulties in interpreting individual differences at the group level (e.g. it is not clear which particular processes are contributing to any observed group differences[19]), we did not analyze reaction times in Experiments 1 and 2. Since participants completed *both* the Self and Other conditions in Experiments 3-6, we were able to directly compare the speed in which choices were made for the self compared to those made for others. Because we did not limit participants' decision time, reaction times were right-skewed. To help normalize the data for subsequent analyses, we log-transformed (base 10) all reaction times.

First, we expected that the severity of the fairness violation would affect the speed at which choices were made. In line with this, we found a main effect of offer type, such that as Player A's offer became increasingly unfair, participants responded faster (repeated measures ANOVA $F(3,1308)=85.2$, $p<0.001$ (N=437), Fig S5A). Second, we also expected to see a difference in response times for choices made for the self compared to those made for others. It is possible that decisions involving personal benefit or loss (Self condition) are associated with greater automaticity, and thus are made more quickly than those made on behalf of another. It is also possible, however, that choices made for the self are more personally consequential, requiring greater deliberation and reflection, and are thus made more slowly than the non-consequential choices made for others. Analysis revealed that participants were quicker to decide for another (1.89s SD±.56) than for themselves (1.99s SD±.54; ($F(1,436)=33.87$, $p<0.001$, reaction times broken down by offer type and condition: Fig S5/S6), suggesting that choice for others entail less deliberation compared to choice for the self.

Moreover, there was an interaction such that as the offer became increasingly unfair $\left(\begin{smallmatrix} .90 \\ .10 \end{smallmatrix}\right)$, the difference in speed between choices made in the Self and Other conditions diminished ($F(3,1308)=227.3$, $p<0.001$, partial $\eta^2 = 0.34$; Fig 7A, each offer type significantly differing from its neighbor, Fisher LSD post-hoc tests; Ps < 0.05). Although choices for others were made significantly faster than those for the self, it is possible that choosing to compensate requires greater deliberation than when deciding to punish. Thus, in order to control for response type, we directly compared whether retributive choices for others were also made more quickly than retributive choices for the self. In line with this, we found that decisions to 'reverse' the payouts on behalf of another were made significantly faster (1.99s SD±.75) than the same decision for the self (2.11s SD±.69: $t(474)=2.56$, $p=0.01$, Fig S7B). Participants were slower to punish the transgressor after directly experiencing a fairness violation.

Countering the classic notion that third-parties—e.g. juries—respond in a more reflective, deliberative manner, this data suggests that endorsing punishment on behalf of another is actually associated with a faster, more automatic process, compared to when personally responding to a fairness violation. In other words, despite the conventional wisdom that we are more deliberative and thoughtful when acting on behalf of wronged others[20], instead we find that such choices are *less* deliberative. In addition, that retributive responses were associated with greater automaticity, dovetails with existing work indicating that emotion related processes play a guiding role in driving punishment[21,22].

Caveats

It is possible that some participants believed that the most fiscally beneficial move is for Player A to offer a $\binom{.90}{.10}$ split. If Player B then chooses to 'compensate', both players can maximize their payouts by each making \$.90. In other words, joint payoff is maximized if Player A makes an initial unfair offer, and Player B then chooses to compensate him or herself and not apply any punishment to Player A. From this perspective, the wisest strategic move is for Player A to always offer the most unfair split and anything less than a $\binom{.90}{.10}$ split should be construed as leaving 'money on the table'. If this is indeed a strategy that participants employed while playing the task, then all other offers $\binom{.60}{.40} - \binom{.80}{.20}$ should be punished at a higher rate than a $\binom{.90}{.10}$ split, and participants should not display any punitive behavior when offered a $\binom{.90}{.10}$ split. Contrary to this, participants' responded with increasingly punitive and retaliatory behavior as the offer became increasingly unfair. However, to check whether participants were operating under this assumption, we debriefed participants at the end of the task and asked them to describe their strategies. Participants' comments during debriefing do not suggest that they believed Player A was acting strategically by offering a highly unfair split (see debriefing section below). Given these factors, it is unlikely that the lack of punishment towards Player A can be explained by participants engaging in the task from the perspective that $\binom{.90}{.10}$ is the most strategic, lucrative, and optimal first move.

Experiments 2-6 Strategies

We asked participants to describe in their own words their strategy for when Player A offered a $.60, $.10 split to them, and to another Player B, and also their strategy for when Player A offered a $.90, $.10 split to them, and to another Player B. This allowed us to explore how participants perceived the intentions of Player A, and to comment on their thought process when deciding to reapportion the payouts. Below we provide a representative sample of the participants' comments for highly unfair and relatively fair splits when they were Player B and when they were Player C.

Question: In your words please describe your strategy for a scenario when Player A kept $.90 and offered $.10 to you.

- *"Instead of tumbling into a vindictive wonderland and punishing A severely, I took the opportunity to make major bank while keeping the playing field even"*
- *"I tried to understand the other person's perspective and tried to equalize by giving both of us .90 instead of focusing on the punishment"*
- *"I was slightly offended by this, but rather than punish player A, I thought it would be more civil to cut the sum evenly in half."*
- *"This is totally unfair and I would overturn the decision, but instead of punishing Player A I would allow for both of us to receive $0.90."*
- *"Feel that A should be punished and would want to reverse the roles, however, they have already played, it is better to give everyone equal and higher money than anyone less."*
- *"That was extremely unfair, so I tried to make it more fair - and even."*
- *"Selfishness shouldn't [sic] be rewarded"*
- *"I was interested in teaching by example. Just because someone was unfair to me doesn't mean I had to be unfair back."*
- *"That is very unfair to me -- it's pretty bad -- shame on player A!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!"*

Question: In your words please describe your strategy for a scenario when Player A kept $.90 and offered $.10 to another Player B (when you were Player C).

- *"I wasn't going to sit around and watch inequality happen, so I choose to eradicate A's advantage/privilege and bringing B up to A's level, so there'd be no income gap/power-advantage. It adds more throughput in the economy, and when I'm B I'd find that a pleasant surprise"*
- *"Its not up to me to forgive player A"*
- *"That was unfair, and I wanted to reverse it so the other player got the unfair payment."*
- *"Player A is greedy and deserves to be punished by only getting $.10 and Player B receiving $.90"*
- *"As I don't want player B to be upset, (especially with ME, since I have the power to change things) I'd upset both people as little as possible"*
- *"That's being greedy and unfair. When I had the chance to right what I felt was wrong, I did it"*
- *"I punished Player A for being selfish, and rewarded Player B because B almost was taken advantage of"*
- *"Grossly unfair. Punish player A if possible, get the most for player B as a higher priority."*

- *"I am still a utilitarian, though it feels more right in this circumstance to reverse the funds for each player with A getting the dime he would have given to B."*
- *"Since player A was being very unfair, I wanted to punish him to make sure he got as little as possible."*

Question: In your words please describe your strategy for a scenario when Player A kept $.60 and offered $.40 to you.

- *"I would choose the option that made both of us get $.60. I felt his offer was somewhat fair so I decided not to deduct anything from Player A."*
- *"I thought it was rather fair, and I don't think I penalized anyway as a result of tying to be close to evenly fair."*
- *"Since Player A tried to be mostly fair, I wanted to maximize the payoff for both of us."*
- *"That is close to fair so I decided to let the offer stand."*
- *"Mostly fair so didn't [sic] punish A, just raised my own stake to .60, also I thought it was kind of fair but decided to make it more equal."*
- *"Since this isn't horribly unfair, I would prefer to give us both .60 or the .50/.50 split....certainly I would not reverse the winnings or punish Player A by giving us both .40."*
- *"This was so close, that it wasn't worth quibbling over 10 cents difference."*
- *"It seemed fair enough. I would have done the same."*

Question: In your words please describe your strategy for a scenario when Player A kept $.60 and offered $.40 to another Player B (when you were Player C).

- *"I thought it was a fair enough offer, although [sic] it could be a little more balanced."*
- *"Once again as long as player A was trying to be fair, then I wanted to try to maximize the payoffs for both players."*
- *"It was somewhat fair, but 50/50 is a better response."*
- *"I maximized the money each player made, as long as it was equal."*
- *"I considered that fair so I made each of them get $.60."*
- *"In this case, I am more likely to give the .50/.50 split because the .10 loss to Player A is still a signal that fairness should be key...however, the original split isn't so unfair that I would penalize A."*
- *"This was fairly equitable, so I would choose to boost B rather than punish A."*
- *"I see that Player A was trying to be reasonably fair, and bump Player B to 60 cents also, in order for both the players to win."*
- *"He thinks he can pull the fleece over b's eyes! He's got something else coming [sic]!"*

Participants' comments indicate that when Player A offered a $.90, $.10 split, participants genuinely felt that it was unfair and not a strategic first move. In fact, none of the 898 participants indicated that a $.90, $.10 split was an optimal first move that could maximize all Players' fiscal payout. Given this, we are confident that participants were not interpreting Player A's highly unfair offers as an intention to be cooperative by maximizing the Players' payouts.

SUPPLEMENTARY METHODS

Experiment 1 Protocol

At the start of each trial in Experiment 1 neither Player A nor Player B knew which options would be made available to Player B on that trial. Randomly pairing the options on each trial such that the option to compensate was not always available prevents Player A from believing that a $9/1 offer is the most optimal and beneficial first move for both Player A and Player B. That is, a $9/1 split can only be considered optimal if Player A knows that Player B has the option to compensate. With this framework, Player A cannot rely on a strategy that offering a $9/1 split maximizes both participants' payouts. Additionally, this dynamic simulates a more naturalistic setting, where people in real world situations typically do not have full information on how others will respond to their choices.

Participants were also told that one trial would be randomly selected by the computer to be paid out. Half the time the trial would be paid out according to the decision of Player B on that trial, and half of the time the computer would treat the trial like a dictator game such that the randomly selected trial would be paid out according to the split suggested by Player A. This payout structure was added so that Player B would know that 50% of the time Player A could maximize their own payout irrespective of Player B's decisions, and to minimize fair offers from Player As. Given that 50% of the trials would be paid out as dictator games, Player As should employ a strategy that will maximize their payouts (a selfish strategy). In addition to the $10 show up fee, participants were able to make an additional payout based on their and their partners' choices (up to $9). Finally, participants were told that during a given experimental session, they would play against many other players in the room, and that on each round (70 rounds in total) they would be paired with a different partner, therefore they should treat each round as a new interaction.

The experimenter read the following instructions out loud to all participants:

"Today you are going to be being playing a game with other players in the room. You will be playing for real money and you will be paid out based on your decisions and the decisions of others. In this game there are two players – Players A and B. At the start of each round Player A will be endowed with $10 and will decide how to divide the $10 between themselves and Player B. For example, Player A can divide the money so that he/she gets $9 and Player B gets $1. Player As can offer however much money they want to Player B's so long as it is in whole dollar increments between $1 and $9. Player A will keep the remaining amount. That is, if Player A offers B $1, they retain $9 for themselves. After Player A has made an offer to Player B, Player B will then be presented with options to reapportion the money. Altogether there are five types of options in this game, however, it is important to note that only 2 of these 5 options will be available in any single given interaction."

"Lets say that Player A divides the $10 by keeping $8 and offering Player B $2. Given A's division of the money, here are the five types of options that B could have. The first option would allow B to decrease A's monetary outcome such that both players receive $2; the second option would increase B's monetary outcome such that both players receive $8; the third option would equally distribute the money between A and B such that both players receive $5; the fourth option would reverse the offer from A such that A will receive $2 and B will receive $8; and the fifth option would accept the offer from A, without changing it. Remember: on any given trial, B's will only have two of these five option types available to them. Only 2 options will be presented at one time to Player B. These two options could be any combination of the options described earlier. The available option types are randomly selected from trial to trial."

"To determine the final payouts for all Players, the program will randomly select 1 trial at the end of the experiment. This one trial will be realized—that is, paid out. 50% of the time both players will be paid whatever B decided on that trial. 50% of the time the computer will ignore B's choice, and simply apportion the money as A had proposed. This means that half of the time, whatever A decided is what happens (like a "dictator game").

Experiments 2-6 Protocol

Amazon Mechanical Turk

Participants were recruited for these experiments using the online labour market Amazon Mechanical Turk (AMT). AMT is an online market in which "employers" can pay "workers" to complete relatively short tasks for small amounts of money. In our experiments, our participants ("workers") received a baseline non-waivable payment of $0.50, in addition to which they could receive a bonus depending on their choices. In other words, participants were incentivized to report their real preferences as one of their choices would be realized and paid out.

One benefit of AMT is that it provides a subject pool that is typically much more diverse than the subject pools available at most American universities [11]—including variation across age, ethnicity, and socio-economic status—ultimately providing a more representative sample of the true population. In an initial pilot study we recruited participants from around the world. However, we discovered through the online debriefing portion of the experiment—where participants were asked to write down their choice strategies—that task comprehension was often poor. To ensure a high level of data quality (e.g. from participants who completely understood the task), we decided to restrict our recruitment to participants based in the United States.

The use of AMT presents some potential concerns not otherwise present in laboratory settings. To address these concerns, a number of studies have explored the validity of data gathered on AMT. Across multiple domains, the behavior reported from AMT participants parallels the behavior found in laboratory participants, indicating the validity and reliability of AMT data [11-17]. In fact, even economic games run on AMT that use stakes 10-fold lower than those run within the laboratory demonstrate similar behavioral results [13,14].

Amazon Mechanical Turk Procedure

While each of the five experiments was slightly different (see below), all the experiments began with a similar set of instructions. When explaining the rules of the game, the instructions explicitly framed offers as fair and unfair. This was done for two reasons. First, in order to make sure that online Mturkers were aware of what a fairness violation was, and second to minimize how participants interpreted the offers.

Instructions for Experiments 2-6

"The purpose of this task is to study how people make decisions. You will be making decisions that affect the monetary outcomes of YOURSELF and OTHERS. You will be playing multiple rounds of a game. Each round will be one of two scenarios. You will be informed of which scenario you are playing at the start of each round. There are two scenarios: in scenario 1 you will be playing as player B and in scenario 2 you will be playing as player C" (a figure was shown illustrating the dynamics of the game).

"In both scenarios, Player A has been allotted $1.00. You will interact with a different Player A on each round. Each Player A has already decided how much of their $1 to share with Player B. Player A can decide to split the $1 however they want, ranging from keeping nearly all the money ($.90 for themselves and $.10 for their partner) or splitting the money evenly ($.50 for themselves and $.50 for their partner). After observing the split that Player A has made, you will be asked to make a decision that will determine the monetary outcome of both Player A and Player B. You can decide to: 1. Decrease Player A's money (thereby punishing them for an unfair offer) 2. Increase Player B's money (thereby compensating them for receiving an unfair offer) 3.

Keep both Players' money the same (thereby accepting the offer from Player A) 4. Reverse both Players' money (thereby ensuring that Player A is punished and Player B is compensated) 5. Equally split the money between both Players. Ultimately, you will decide how much money Player A and B actually receive. "

"IMPORTANT: You will be playing multiple rounds of this game. Sometimes as Player B and sometimes as Player C. In Scenario 1, YOU will be Player B and you are making the choice for <u>your own</u> monetary outcome. In other words, you will have a personal stake in the outcomes, and you will have the chance to make additional money depending on your choices. In Scenario 2, you are making the choice on <u>behalf of a 3rd person</u>, another Player B. That is, you will not yourself be invested in the decision when you are deciding as Player C, but you will make choices that will effect the monetary outcomes of another Player B. When you are making decisions as Player C, you will not make an additional bonus but Players A and B could make additional money depending on your choices."

"During the task itself, please place your hands on the keys S, D, F, H, and J. Each of these keys will correspond to a different response. Once you hit the key, your decision will be recorded and the next trial will appear, so please be certain of your choice before hitting the key. YOUR MOUSE WILL NOT WORK DURING THE TASK so do not use the mouse to tick the boxes."

Participants were then presented with an example trial, and then explained step-by-step what happened in the example trial, Fig S4.

*"In this example, Player A unfairly divides the $1.00 endowment by keeping $.80 and giving Player B $.20. Immediately below that, you can see the options that will change the monetary outcomes of both players. Hitting the **S** key will result in reversing the outcomes of Player A and Player B (Player A will be punished for offering an unfair division and Player B will be compensated for receiving an unfair offer). Hitting the **D** key will result in an equal split between the players. Hitting the **F** key will result in decreasing Player A's outcome while keeping Player B's outcome the same (Player A will be punished for offering an unfair division and will thus receive $.20, the same amount that Player B will receive, $.20). Hitting the **H** key will result in increasing Player B's outcome, while keeping Player A's outcome the same (both Players will receive $.80; thus, Player B is being compensated for receiving an unfair division). Hitting the **J** key will result in keeping the Players' outcomes the same as suggested by Player A. IMPORTANT: The choices that you will see will never be in the same order, so pay attention!"*

SUPPLEMENTARY REFERENCES

1 Von Neumann, J. & Morgenstern, O. *Theory of games and economic behavior*. (Princeton university press, 1944).

2 Fehr, E. & Fischbacher, U. Why social preferences matter - The impact of non-selfish motives on competition, cooperation and incentives. *Econ J* **112**, C1-C33, doi:Doi 10.1111/1468-0297.00027 (2002).

3 Fehr, E. & Schmidt, K. M. A theory of fairness, competition, and cooperation. *Q J Econ* **114**, 817-868, doi:Doi 10.1162/003355399556151 (1999).

4 Blount, S. When Social Outcomes Arent Fair - the Effect of Causal Attributions on Preferences. *Organ Behav Hum Dec* **63**, 131-144, doi:Doi 10.1006/Obhd.1995.1068 (1995).

5 Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R. & Smirnov, O. Egalitarian motives in humans. *Nature* **446**, 794-796, doi:Doi 10.1038/Nature05651 (2007).

6 Camerer, C. *Behavioral game theory : experiments in strategic interaction*. (Russell Sage Foundation; Princeton University Press, 2003).

7 Guth, W., Schmittberger, R. & Schwarze, B. An Experimental-Analysis of Ultimatum Bargaining. *J Econ Behav Organ* **3**, 367-388, doi:Doi 10.1016/0167-2681(82)90011-7 (1982).

8 Carlsmith, K. M., Darley, J. M. & Robinson, P. H. Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology* **83**, 284-299, doi:Doi 10.1037//0022-3514.83.2.284 (2002).

9 Weitekamp, E. Reparative Justice. *European Journal of Criminal Policy and Research* **1**, 70-93 (1993).

10 Charness, G. & Rabin, M. Understanding social preferences with simple tests. *Q J Econ* **117**, 817-869, doi:Doi 10.1162/003355302760193904 (2002).

11 Buhrmester, M., Kwang, T. & Gosling, S. D. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspect Psychol Sci* **6**, 3-5, doi:Doi 10.1177/1745691610393980 (2011).

12 Rand, D. G. The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *J Theor Biol* **299**, 172-179, doi:Doi 10.1016/J.Jtbi.2011.03.004 (2012).

13 Horton, J. J., Rand, D. G. & Zeckhauser, R. J. The online laboratory: conducting experiments in a real labor market. *Exp Econ* **14**, 399-425, doi:Doi 10.1007/S10683-011-9273-9 (2011).

14 Suri, S. & Watts, D. J. Cooperation and Contagion in Web-Based, Networked Public Goods Experiments. *PloS one* **6**, doi:ARTN e16836 DOI 10.1371/journal.pone.0016836 (2011).

15 Paolacci, G., Chandler, J. & Ipeirotis, P. G. Running experiments on Amazon Mechanical Turk. *Judgm Decis Mak* **5**, 411-419 (2010).

16 Mason, W. S., S. Conducting behavioral research on Amazon's Mechanical Turk *Behavioral Research Methods* **44**, 1-23 (2012).

17 Goodman, J. K., Cryder, C. E. & Cheema, A. Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. *J Behav Decis Making* **26**, 213-224, doi:Doi 10.1002/Bdm.1753 (2013).

18 Schimmack, U. The Ironic Effect of Significant Results on the Credibility of Multiple-Study Articles. *Psychol Methods* **17**, 551-566, doi:Doi 10.1037/A0029487 (2012).

19 Salthouse, T. A. & Hedden, T. Interpreting reaction time measures in between-group comparisons. *J Clin Exp Neuropsychol* **24**, 858-872, doi:10.1076/jcen.24.7.858.8392 (2002).

20 Fehr, E. & Fischbacher, U. Third-party punishment and social norms. *Evol Hum Behav* **25**, 63-87, doi:Doi 10.1016/S1090-5138(04)00005-4 (2004).

21 Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E. & Cohen, J. D. The neural basis of economic decision-making in the Ultimatum Game. *Science* **300**, 1755-1758 (2003).

22 Fehr, E. & Gachter, S. Altruistic punishment in humans. *Nature* **415**, 137-140, doi:Doi 10.1038/415137a (2002).