LETTER

doi:10.1038/nature16981

Third-party punishment as a costly signal of trustworthiness

Jillian J. Jordan[1], Moshe Hoffman[2], Paul Bloom[1] & David G. Rand[1,3,4]

Third-party punishment (TPP)[1–7], in which unaffected observers punish selfishness, promotes cooperation by deterring defection. But why should individuals choose to bear the costs of punishing? We present a game theoretic model of TPP as a costly signal[8–10] of trustworthiness. Our model is based on individual differences in the costs and/or benefits of being trustworthy. We argue that individuals for whom trustworthiness is payoff-maximizing will find TPP to be less net costly (for example, because mechanisms[11] that incentivize some individuals to be trustworthy also create benefits for deterring selfishness via TPP). We show that because of this relationship, it can be advantageous for individuals to punish selfishness in order to signal that they are not selfish themselves. We then empirically validate our model using economic game experiments. We show that TPP is indeed a signal of trustworthiness: third-party punishers are trusted more, and actually behave in a more trustworthy way, than non-punishers. Furthermore, as predicted by our model, introducing a more informative signal—the opportunity to help directly—attenuates these signalling effects. When potential punishers have the chance to help, they are less likely to punish, and punishment is perceived as, and actually is, a weaker signal of trustworthiness. Costly helping, in contrast, is a strong and highly used signal even when TPP is also possible. Together, our model and experiments provide a formal reputational account of TPP, and demonstrate how the costs of punishing may be recouped by the long-run benefits of signalling one's trustworthiness.

Costly third-party punishment (TPP) is widely observed in laboratory[1–4,7] and field[5,6] experiments (although see ref. 12), and appears to be universal across cultures[13]. While collectively beneficial, TPP poses a puzzle: why should individuals incur the costs of punishment?

We propose an answer based on reputation[4,14–18]. Specifically, we introduce a game theoretic model of TPP as a costly signal of trustworthiness: if you see me punish selfishness, it can signal that I will not be selfish to you. Our model involves a partner-choice[19] game with two roles. In each interaction, the 'Signaller' decides whether to send one or more costly signals; then the 'Chooser' decides whether to partner with the Signaller.

As with all costly signalling models[8–10], our model is based on individual differences: two 'types' of Signallers differ in their quality as interaction partners. For trustworthy types, it is payoff-maximizing to cooperate when trusted; for exploitative types, it is payoff-maximizing to defect. Choosers thus benefit from partnering with trustworthy Signallers, but are harmed by partnering with exploitative Signallers.

Signallers' types are fixed, but not directly observable. Therefore, Choosers must base their partner choice on the aforementioned costly signals. In each interaction, the Signaller's cost of signalling is either small (less than the benefit of being chosen as a partner) or large (greater than the benefit of being chosen). It is thus beneficial to signal (in order to be chosen) when the cost is small, but not large. The key premise of costly signalling is that high-quality types are more likely to experience small signalling costs than low-quality types (and are thus more likely to signal). Therefore, signals convey information about Signallers' types, and Choosers benefit from preferring partners who signal.

How does this relate to TPP and trustworthiness? We argue that TPP will typically be less net costly for trustworthy types (that is, individuals who find it payoff-maximizing to cooperate when trusted). Because TPP deters future harm against others, punishing may benefit the punisher (for example, via direct reciprocity from the victim of the punished transgression, or rewards from institutions or leaders seeking to promote cooperation); and these benefits should be larger for trustworthy types, because the same mechanisms[11] that make trustworthy behaviour advantageous also increase the benefits of preventing harm against others. (This argument implies that the costly signalling mechanism we propose may interact positively with other mechanisms for TPP that are based on deterrence benefits.) Furthermore, because trustworthy types are more desirable to interact with than exploitative types, they typically attract more partners—which may reduce TPP costs by offering protection against retaliation and facilitating coordinated punishment[20]. See Supplementary Information sections 1.2.3 and 1.3 and Extended Data Fig. 1 for formal models of these two microfoundations for our central argument.

When TPP is less net costly for trustworthy types, it can serve as a costly signal of trustworthiness. Agents should thus sometimes punish for the express purpose of signalling their trustworthiness to Choosers (like a peacock's tail signals genetic quality)—specifically, when the deterrence benefits of TPP are too small to outweigh the costs on their own (otherwise, TPP would occur without signalling), but the reputational benefit of appearing trustworthy makes TPP worthwhile.

Although TPP can convey information about type, there are often several possible ways to signal trustworthiness, and TPP is not always the most informative. Therefore, a crucial prediction of this signalling account is that when a more informative signal is available, the signalling value of TPP should be attenuated and less TPP should occur. To illustrate this fact, our model also includes the possibility of signalling via costly helping of third parties: because being trustworthy and helping both involve paying costs to benefit others, helping should typically be a very informative signal of trustworthiness (see Supplementary Information section 1.2.3).

Agents in our model make decisions across three different scenarios in which Signallers have the opportunity to engage in (1) TPP, (2) third-party helping, or (3) both. In each scenario, Choosers know which signals were available to Signallers. An agent's strategy specifies her actions as both the Signaller and Chooser in each scenario.

Our equilibrium analysis identifies Nash equilibria that are robust against indirect invasion (RAII)[21] (and thus likely to be favoured by natural selection; see Supplementary Information section 2.1). We also directly test which strategies are favoured by selection using stochastic evolutionary dynamics where agents interact at random to earn payoffs, strategies with higher payoffs become more common, and mutation

[1]Department of Psychology, Yale University, New Haven, Connecticut 06511, USA. [2]Program for Evolutionary Dynamics, Harvard University, Cambridge, Massachusetts 02138, USA. [3]Department of Economics, Yale University, New Haven, Connecticut 06511, USA. [4]School of Management, Yale University, New Haven, Connecticut 06511, USA.

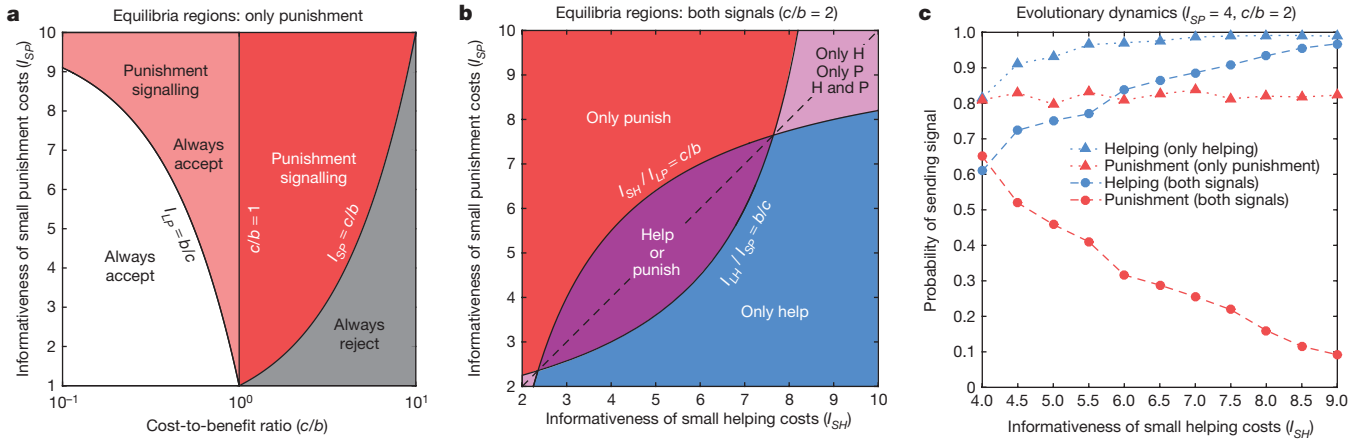


Figure 1 | A model of TPP as a costly signal of trustworthiness. Shown are the results of equilibrium calculations (in which we identify Nash equilibria that are RAII) and evolutionary dynamics, where I_{SP} (I_{SH}) is the informativeness of small punishing (helping) costs (that is, the ratio between the probabilities that trustworthy versus exploitative Signallers experience small punishment (helping) costs), I_{LP} (I_{LH}) is the informativeness of large punishing (helping) costs (the inverse ratio, using large costs instead of small), b is the expected benefit of partner choice (product of the probability a Signaller is trustworthy and the gain to Choosers of accepting trustworthy Signallers), and c is the expected cost of partner choice (the same product, but for exploitative Signallers). **a**, Equilibria when only punishment is possible. 'Punishment-signalling', in which Signallers punish when their punishing costs are small and Choosers only accept punishers, is an equilibrium in red and pink, when (i) large punishment costs are informative enough that the expected payoff of accepting a Signaller with a large cost is negative ($I_{LP} > b/c$); and (ii) small punishment costs are informative enough that the expected payoff of accepting a Signaller with a small cost is positive ($I_{SP} > c/b$). It is the unique equilibrium in red when also (iii) the expected payoff of accepting a random Signaller is negative ($c/b > 1$). We also see pooling equilibria in which Signallers never punish and Choosers always accept (in white and pink) or reject (in grey). For visualization, we fix the probability that an exploitative Signaller has a small punishment cost at 0.1, and vary the probability that a trustworthy Signaller has a small punishment cost (showing the resulting I_{SP}). **b**, Equilibria when both signals are possible. We consider parameters where punishment-signalling and helping-signalling are the unique equilibria in scenarios 1 and 2, respectively (fixing $c/b = 2$, and showing I_{SP} and I_{SH} values >2). In scenario 3, an only-helping strategy profile, in which Signallers help when their helping costs are small but never punish, and Choosers demand helping but ignore punishment, is an equilibrium in blue and light purple, when (i) large helping costs are sufficiently more informative than small punishment costs that the expected payoff of accepting any Signaller with a large helping cost is negative, even if she also has a small punishing cost ($I_{LH} / I_{SP} > b/c$). It is the unique equilibrium in blue, when also (ii) small helping costs are sufficiently more informative than large punishment costs that the expected payoff of accepting a Signaller with a small helping cost remains positive even if she also has a large punishing cost ($I_{SH} / I_{LP} > c/b$). When helping and punishment are similarly informative (around the diagonal), we also see equilibria in which Choosers treat them equally, demanding helping or punishment (dark purple), or helping and punishment (light purple). For visualization, we vary the informativeness of both punishment and helping as in **a**, showing the resulting I_{SP} and I_{SH} values. **c**, Evolutionary dynamics when punishment is moderately informative and helping is increasingly more informative. Here we use agent-based simulations and fix $I_{SP} = 4$, $I_{LP} = 1.5$ and $c/b = 2$, such that when punishment is the only signal (red triangles), it is favoured by evolution (regardless of how informative helping is, because helping is not possible in that scenario). We then vary the informativeness of helping as in **b**. We see that when both signals are available, evolution increasingly disfavours punishment as helping becomes increasingly informative (red circles). In contrast, Signallers help at high rates regardless of whether helping is the only signal (blue triangles) or both signals are available (blue circles). We plot Signallers' probabilities of sending each signal when the cost is low; for scenario 3, we plot the average of the probabilities of punishing (helping) when (i) only the punishing (helping) cost is low, and (ii) both costs are low. See Extended Data Fig. 2 for Chooser strategies and disaggregated Signaller strategies.

maintains variation (see Supplementary Information section 3.1). This process can describe genetic evolution, as well as social learning whereby people imitate successful others.

We first consider scenario 1, where Signallers have the opportunity to punish but not help. Here a punishment-signalling strategy profile (in which Signallers punish when experiencing small signalling costs, and Choosers only accept Signallers who punish) can be an equilibrium when punishment is sufficiently informative: that is, when trustworthy types are sufficiently more likely to receive small punishment costs, and less likely to receive large punishment costs, than exploitative types (see Fig. 1a for precise conditions). Thus, we confirm that TPP can signal trustworthiness when it is the only available signal. By symmetry, the same is true for helping when it is the only available signal (scenario 2). See Supplementary Information section 2.2 for details.

What, then, happens in scenario 3 when TPP and helping are both possible? If helping is more informative, TPP may be ignored. To see why, consider a Signaller who punishes but does not help. If she did not have the opportunity to help, her choice to punish conveys positive information, and a Chooser might accept her. However, if helping was possible, her choice not to help conveys negative information—and when not helping is informative enough to outweigh the positive effect of punishing, the same Chooser might reject her.

To formalize this argument, we vary the informativeness of TPP and helping in scenario 3. We focus on the parameter region where both TPP and helping are informative enough to serve as signals on their own (that is, punishment-signalling and helping-signalling are the unique equilibria in scenarios 1 and 2, respectively; everywhere in Fig. 1b). We find that when the informativeness of the two signals is sufficiently similar, there are equilibria in which Signallers are equally likely to engage in TPP and helping, and Choosers equally demand TPP and helping. However, as the informativeness of helping increases, and/or the informativeness of TPP decreases, the unique equilibrium becomes an only-helping strategy profile in which helping is signalled and demanded, and TPP is ignored. Specifically, only-helping becomes the unique equilibrium when Choosers receive (i) a positive expected payoff from accepting any Signaller with a small helping cost (even if she has a large punishing cost), and (ii) a negative expected payoff from accepting any Signaller with a large helping cost (even if she has a small punishing cost). See Fig. 1b for precise conditions.

Critically, then, there are parameter regions in which it is an equilibrium to punish (and condition partner choice on punishment) in scenario 1 but not in scenario 3. Evolutionary dynamics show that as a result, TPP can evolve as a costly signal that is preferentially used when helping is not possible (Fig. 1c and Extended Data Fig. 2).

Our model thus makes clear predictions. First, when TPP is the only possible signal, it should be perceived as, and should actually be, an honest signal of trustworthiness. Second, when a more informative signal (for example, helping) is also available, third parties should be less likely to punish, and the perceived and actual signalling value of TPP should be attenuated. Third, the same should not be true of helping, which should continue to serve as a strong signal even when TPP is possible.

We next test these predictions in a two-stage economic game conducted using Amazon Mechanical Turk in which TPP and helping signals can be sent, and then partner choice occurs (Extended Data Fig. 3 illustrates the experimental setup, and Supplementary Information section 4 discusses the link between our theoretical and experimental setups). As in our model, there are two roles in this game: Signaller and Chooser.

In the first stage, the Signaller participants in a TPP game[1] (TPPG), interacting with people other than the Chooser. In the TPPG, a Helper decides whether to share money with a Recipient, and an unaffected Punisher decides whether to pay to punish the Helper if the Helper is selfish. To investigate the three scenarios from the model in which helping, punishment or both are available as signals, we manipulate whether the Signaller participates in the TPPG as the Helper, Punisher or both (playing twice with two different sets of other people).

The second stage captures the psychology of partner choice using a trust game (TG). Here, both the Signaller and Chooser participate. The Chooser first decides how much of an endowment to send to the Signaller; any money sent is tripled. The Signaller then decides how much to return to the Chooser. The Chooser can condition her sending on the Signaller's behaviour in the TPPG—and the Signaller knows this when deciding how to behave in the TPPG.

Overall, therefore, our experiment is designed to include opportunities to signal via TPP and/or helping, and to make helping more informative than TPP (see Supplementary Information section 5.1 for further discussion).

The results confirm our theoretical predictions. First, in the punishment-only condition (where punishment is the only available signal, $n = 397$ Signaller–Chooser pairs), punishment is perceived by Choosers as a signal of trustworthiness: Choosers trust Signallers who punish in the TPPG more than those who do not (sending 16 percentage points more to punishers than non-punishers, $P < 0.001$, Fig. 2a). Furthermore, punishment actually is an honest signal of trustworthiness: Signallers who punish return significantly more in the TG than non-punishers (returning 8 percentage points more, $P = 0.001$, Fig. 2b). P values generated using linear regression with robust standard errors; see Supplementary Information section 5.2.

Second, in the punishment-plus-helping condition (where helping is also possible, $n = 393$ Signaller–Chooser pairs), Signallers use punishment less than in the punishment-only condition: only 30% of Signallers punish in punishment-plus-helping, compared to 41% in punishment-only ($P = 0.002$, Fig. 2c). Furthermore, providing the option to help attenuates the perceived and actual signalling value of punishment: in the punishment-plus-helping condition, controlling for helping, Choosers only trust punishers slightly more than non-punishers (4 percentage points more sent to punishers than non-punishers, $P = 0.004$, Fig. 2a), and Signallers who punish in the TPPG do not return significantly more in the TG than non-punishers (0.3 percentage points less returned by punishers than non-punishers, $P = 0.900$, Fig. 2b). Thus the effects of punishment on trust and trustworthiness are significantly smaller in punishment-plus-helping than punishment-only (interactions: $P < 0.001$ and $P = 0.016$, respectively). See Supplementary Information section 5.2 for details.

Third, in the helping-only condition ($n = 409$ Signaller–Chooser pairs), just as many Signallers help (81%) as in the punishment-plus-helping condition (82%) ($P = 0.650$, Fig. 2c). Furthermore, in both conditions, Choosers preferentially trust Signallers who help (39 percentage points more sent to helpers in helping-only,

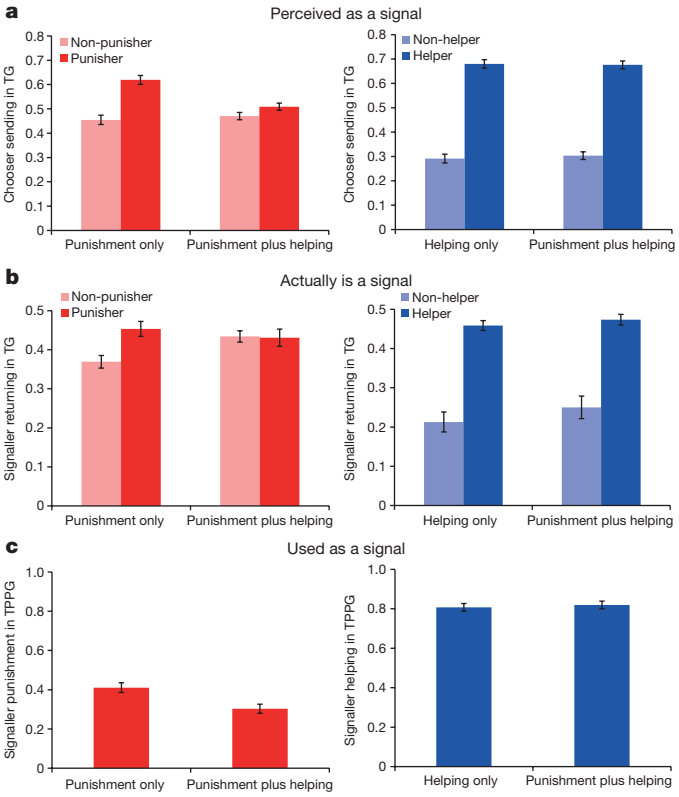


Figure 2 | Behavioural experiments confirm key predictions of our model. a, TPP is perceived as a signal of trustworthiness. In the TG, Choosers trust Signallers who punished in the TPPG more than non-punishers. This effect is much larger, however, when punishment is the only available signal (punishment-only) than when both punishment and helping are possible (punishment-plus-helping) (left column). In contrast, Choosers trust Signallers who helped in the TPPG much more than non-helpers, regardless of whether punishment is possible (right column). **b**, TPP actually is a signal of trustworthiness. Signallers who punished in the TPPG return more in the TG than non-punishers, but this effect is much larger when Signallers did not have the chance to help (punishment-only) than when they did (punishment-plus-helping) (left column). In contrast, Signallers who helped in the TPPG return more than non-helpers, regardless of whether they had the opportunity to punish (right column). **c**, TPP is used as a signal of trustworthiness. Signallers are more likely to punish when they do not have the opportunity to help (punishment-only) than when they do (punishment-plus-helping) (left column). In contrast, they are just as likely to help when they do not have the opportunity to punish (helping-only) as when they do (punishment-plus-helping) (right column). **a** and **b** show predicted mean trust and trustworthiness, respectively (in punishment-plus-helping, regression models include terms for both punishment and helping, but not their interaction, as there were no significant punishment-by-helping interactions); **c** shows predicted probability of sending the relevant signal from logistic regressions including condition. Error bars are ± 1 s.e.m.

$P < 0.001$; 37 percentage points more sent in punishment-plus-helping (controlling for TPP), $P < 0.001$, Fig. 2a), and Signallers who help are more trustworthy (25 percentage points more returned by helpers in helping-only, $P < 0.001$; 22 percentage points more returned in punishment-plus-helping (controlling for TPP), $P < 0.001$, Fig. 2b). These differences between conditions are not significant (interactions: $P = 0.539$ and $P = 0.623$, respectively). Thus, while helping attenuates the signalling value of TPP, TPP does not attenuate the signalling value of helping.

These results offer clear support for our model of TPP as a costly signal of trustworthiness. We therefore provide evidence that people may punish to provide information about their character to observers, rather than just to harm defectors or deter selfishness. This theory helps to reconcile conflicting previous experimental results about whether TPP

confers reputational benefits. Our conclusion that the signalling value of TPP is mitigated when more informative signals of trustworthiness are also available explains why a large positive effect of punishment on trust was found in one experiment in which helping information was absent[22], while little effect was found in another experiment in which helping was observable[16]. This conclusion also provides an explanation for why TPP and trustworthiness were found to be uncorrelated in an experiment in which both punishment and helping were possible[23]. Finally, our theory also explains why participants preferred punishers as partners to a greater extent in situations in which participants could benefit from choosing a prosocial partner[24].

Our results cannot be explained by the alternative theory that TPP is perceived as a signal of willingness to retaliate when harmed directly (although TPP may signal retaliation in other contexts), because retaliation is not possible in the TG. Even if Choosers sent more to punishing Signallers out of an 'irrational' fear of retaliation, helping information should not attenuate this effect (as helping is unlikely to be a more informative signal of retaliation than TPP). Furthermore, an additional experiment (Extended Data Fig. 4 and Supplementary Information section 6) finds that TPP elicits larger reputational benefits when stage 2 is a TG than an ultimatum game (where signalling retaliatoriness is advantageous).

Importantly, punishers need not be consciously seeking to signal their trustworthiness; at a proximate level, TPP may be motivated by emotions like moral outrage[1,3]. Thus, TPP may be based on social heuristics[25] rather than explicit reasoning, and is unlikely to be perfectly sensitive to context—signalling motives may 'spill over'[26] to settings where TPP cannot function as a signal (for example, anonymous interactions, or settings in which engaging with trustworthy Signallers is not actually advantageous to Choosers, such as the Dictator Game[27]).

Relatedly, while our model assumes that different types of individuals have different costs of TPP and helping, and different optimal responses to being trusted, our experiments do not vary subjects' payoffs of punishing, helping and being trustworthy. Instead, the experiments tap into participants' pre-existing inclinations to punish, help and reciprocate the trust of others, reflecting the incentives experienced in daily life[25]. Thus, because we do not exactly recreate the model in the laboratory, our results are consistent with the idea that the model operates outside of the laboratory (rather than merely showing that participants can reason strategically about a novel game).

In sum, we help answer a fundamental question regarding human nature: why do humans care about selfish behaviours that do not affect them personally? Although TPP may often appear 'altruistic', we show how punishing can be self-interested in the long-run because of reputational benefits. Sometimes punishing wrongdoers is the best way to show that you care.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 November 2015; accepted 8 January 2016.

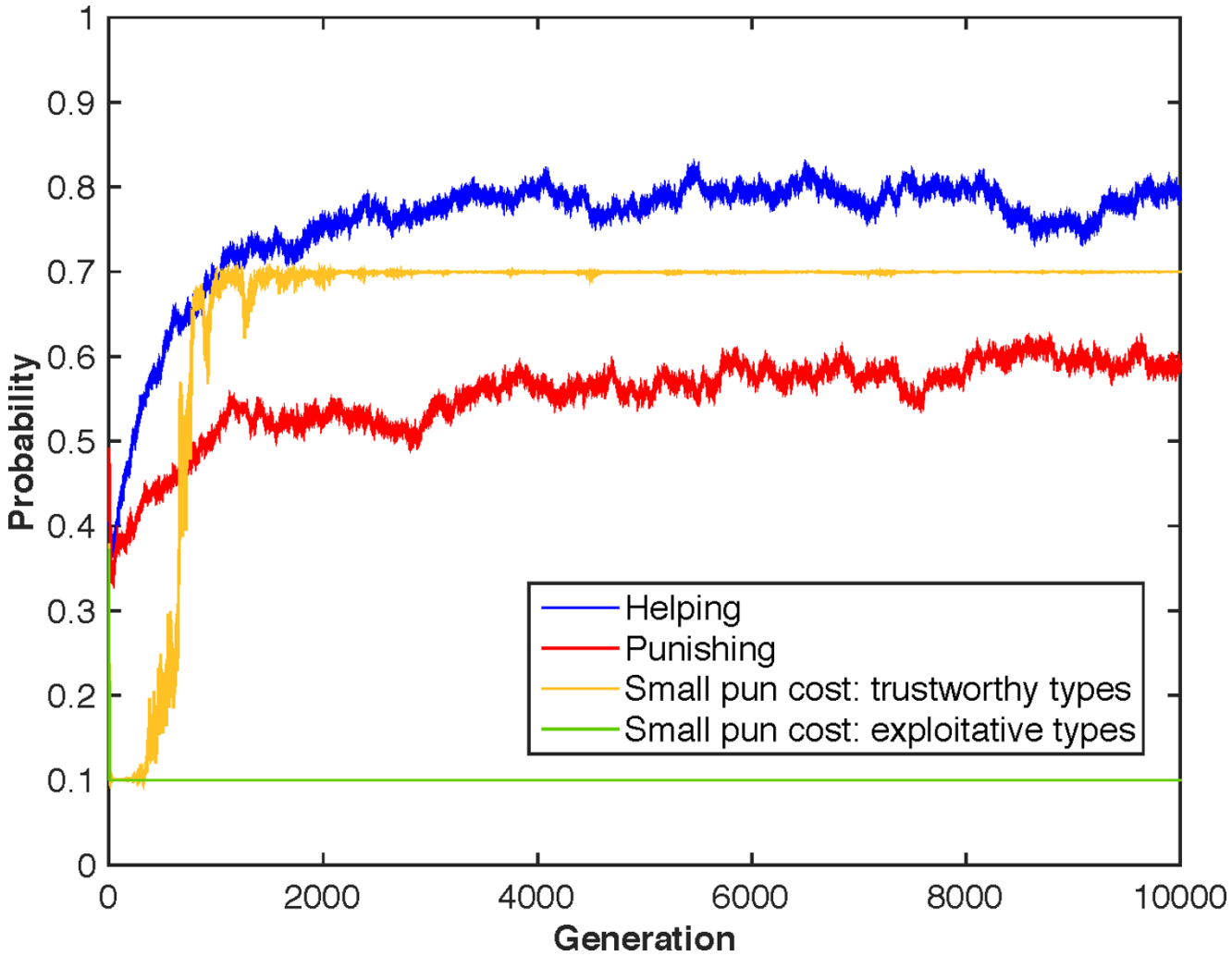
1. Fehr, E. & Fischbacher, U. Third-party punishment and social norms. *Evol. Hum. Behav.* **25,** 63–87 (2004).
2. Goette, L., Huffman, D. & Meier, S. The impact of group membership on cooperation and norm enforcement: evidence using random assignment to real social groups. *Am. Econ. Rev.* **96,** 212–216 (2006).
3. Jordan, J. J., McAuliffe, K. & Rand, D. G. The effects of endowment size and strategy method on third party punishment. *Exp. Econ.* http://dx.doi.org/10.1007/s10683-015-9466-8 (2015).
4. Kurzban, R., DeScioli, P. & O'Brien, E. Audience effects on moralistic punishment. *Evol. Hum. Behav.* **28,** 75–84 (2007).
5. Balafoutas, L. & Nikiforakis, N. Norm enforcement in the city: a natural field experiment. *Eur. Econ. Rev.* **56,** 1773–1785 (2012).
6. Mathew, S. & Boyd, R. Punishment sustains large-scale cooperation in prestate warfare. *Proc. Natl Acad. Sci. USA* **108,** 11375–11380 (2011).
7. FeldmanHall, O., Sokol-Hessner, P., Van Bavel, J. J. & Phelps, E. A. Fairness violations elicit greater punishment on behalf of another than for oneself. *Nature Commun.* **5,** 5306 (2014).
8. Zahavi, A. Mate selection—a selection for a handicap. *J. Theor. Biol.* **53,** 205–214 (1975).
9. Gintis, H., Smith, E. A. & Bowles, S. Costly signaling and cooperation. *J. Theor. Biol.* **213,** 103–119 (2001).
10. Roberts, G. Competitive altruism: from reciprocity to the handicap principle. *Proc. Biol. Sci.* **265,** 427–431 (1998).
11. Rand, D. G. & Nowak, M. Human cooperation. *Trends Cogn. Sci.* **17,** 413–425 (2013).
12. Guala, F. Reciprocity: weak or strong? What punishment experiments do (and do not) demonstrate. *Behav. Brain Sci.* **35,** 1–15 (2012).
13. Henrich, J. *et al.* Costly punishment across human societies. *Science* **312,** 1767–1770 (2006).
14. Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity. *Nature* **437,** 1291–1298 (2005).
15. Raihani, N. J. & Bshary, R. The reputation of punishers. *Trends Ecol. Evol.* **30,** 98–103 (2015).
16. Barclay, P. Reputational benefits for altruistic punishment. *Evol. Hum. Behav.* **27,** 325–344 (2006).
17. Fessler, D. M. & Haley, K. J. in *The Genetic and Cultural Evolution of Cooperation* (ed. Hammerstein, P.) (MIT Press, 2003).
18. Panchanathan, K. & Boyd, R. Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature* **108,** 432–502 (2014).
19. Baumard, N., André, J.-B. & Sperber, D. A mutualistic approach to morality: the evolution of fairness by partner choice. *Behav. Brain Sci.* **36,** 59–78 (2013).
20. Boyd, R., Gintis, H. & Bowles, S. Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science* **328,** 617–620 (2010).
21. van Veelen, M. Robustness against indirect invasions. *Games Econ. Behav.* **74,** 382–393 (2012).
22. Nelissen, R. M. A. The price you pay: cost-dependent reputation effects of altruistic punishment. *Evol. Hum. Behav.* **29,** 242–248 (2008).
23. Peysakhovich, A., Nowak, M. A. & Rand, D. Humans display a 'cooperative phenotype' that is domain general and temporally stable. *Nature Commun.* **5,** 4939 (2014).
24. Horita, Y. Punishers may be chosen as providers but not as recipients. *Lett. Evol. Behav. Sci.* **1,** 6–9 (2010).
25. Bear, A. & Rand, D. G. Intuition, deliberation, and the evolution of cooperation. *Proc. Natl Acad. Sci. USA* **113,** 936–941 (2016).
26. Peysakhovich, P. & Rand, D. G. Habits of virtue: creating norms of cooperation and defection in the laboratory. *Management Science* http://dx.doi.org/10.1287/mnsc.2015.2168 (2015).
27. Raihani, N. J. & Bshary, R. Third-party punishers are rewarded, but third-party helpers even more so. *Evolution* **69,** 993–1003 (2015).

Supplementary Information is available in the online version of the paper.

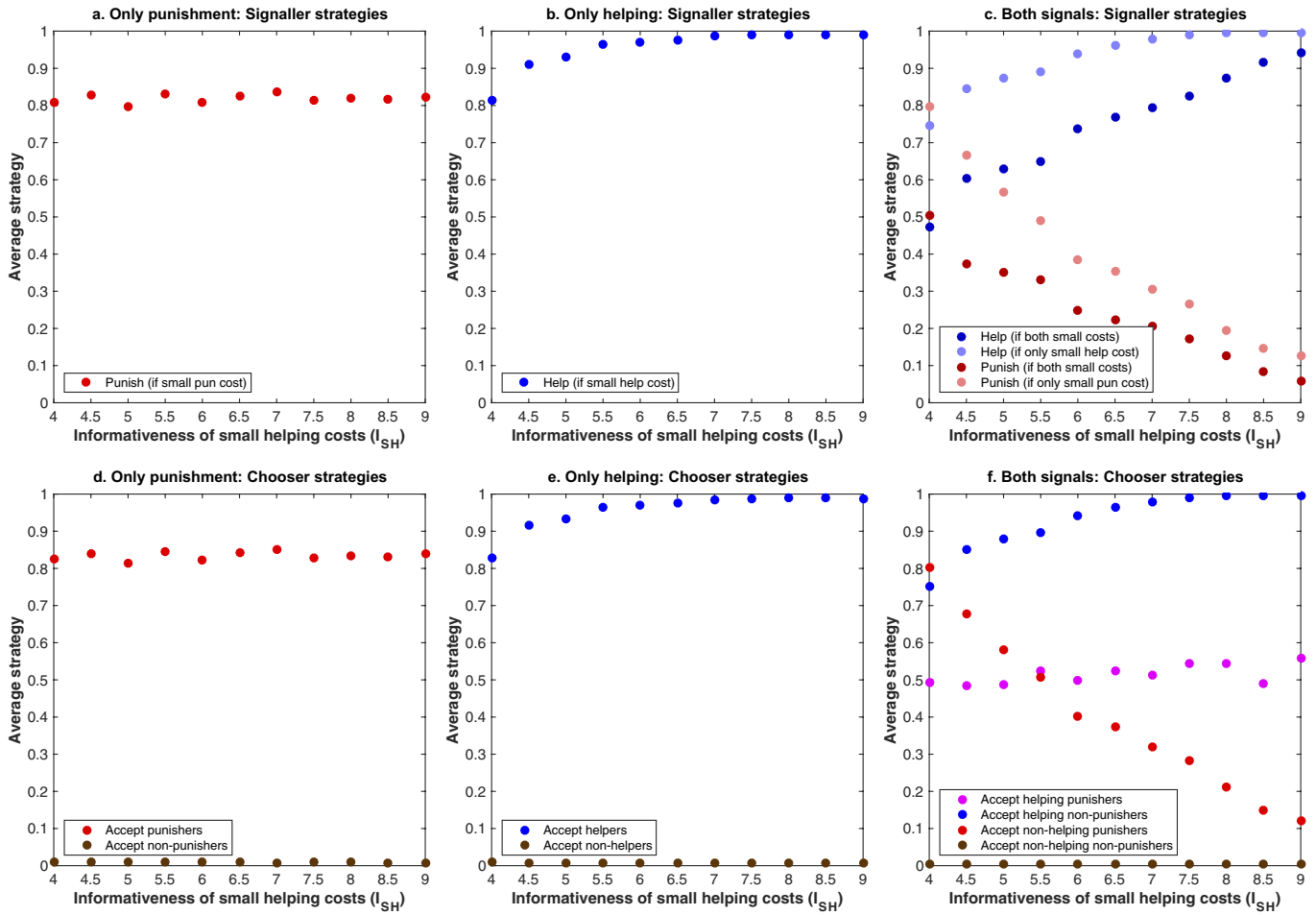
Acknowledgements We gratefully acknowledge the John Templeton Foundation for financial support; A. Bear, R. Boyd, M. Crockett, J. Cone, F. Cushman, E. Fehr, M. Krasnow, R. Kurzban, J. Martin, M. Nowak, N. Raihani, L. Santos, and A. Shaw for helpful feedback; and A. Arechar, Z. Epstein, and G. Kraft-Todd for technical assistance.

Author Contributions J.J.J., M.H. and D.G.R. designed and analysed the model. J.J.J., P.B. and D.G.R. designed the experiments. J.J.J. conducted the experiments and analysed the results. J.J.J., M.H., P.B. and D.G.R. wrote the paper.

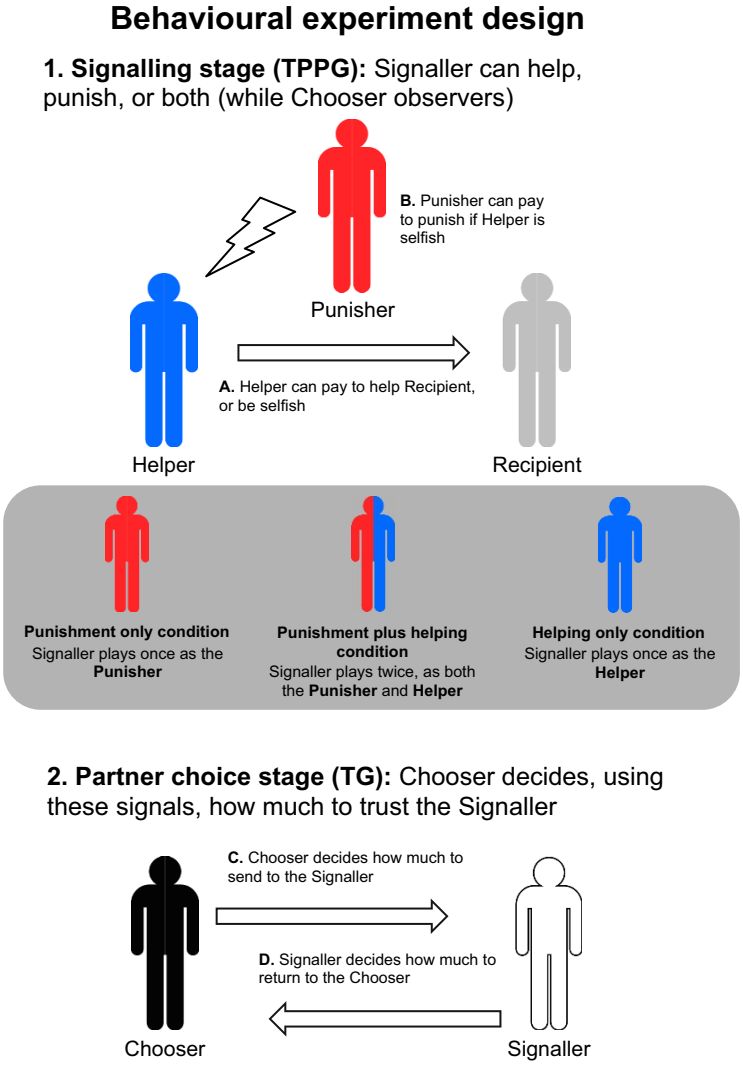
Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.J.J. (jillian.jordan@yale.edu) or D.G.R. (david.rand@yale.edu).



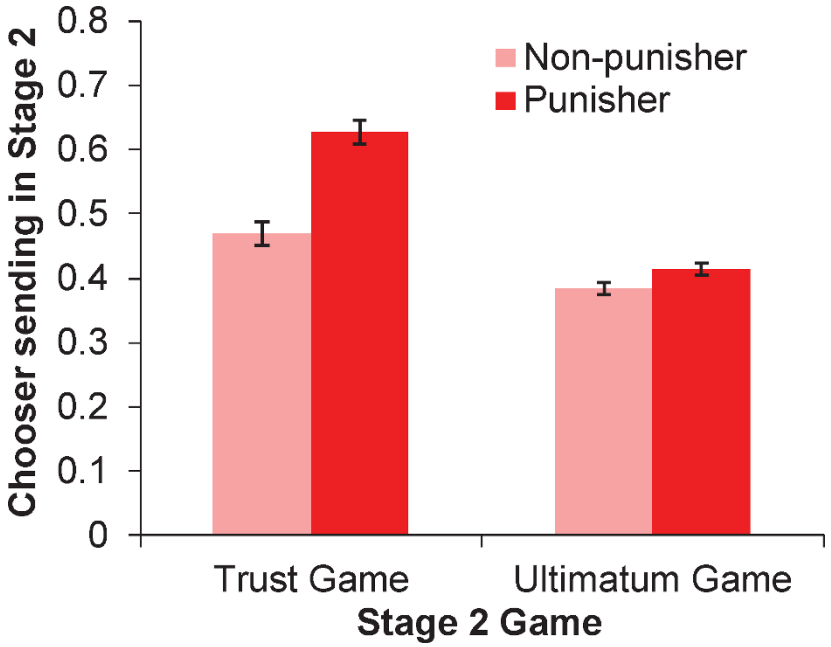
Extended Data Figure 1 | Agent-based simulations from our second microfoundation model in which gaining interaction partners reduces TPP costs. TPP evolves over time in this modified model, in which a Signaller's punishment costs are endogenous (decreasing in the number of times she has been accepted as a partner), rather than exogenously fixed as lower for trustworthy types. We use parameters similar to the main text agent-based simulations, where punishment is moderately informative and helping is more informative. Shown is the average over 500 simulations of Signallers' average probability of helping and punishing (when experiencing the small signalling cost) in each generation, as well as the expected probability of experiencing the small punishing cost for trustworthy and exploitative types (based on the average number of times trustworthy and exploitative types were chosen as partners) at the end of each generation. See Supplementary Information section 1.3.2 for a detailed description of our second microfoundation model.



Extended Data Figure 2 | Full agent-based simulation results from the main text model. Here, we present the Signaller and Chooser strategies for each scenario from our main model agent-based simulations, a summary of which is shown in Fig. 1c. In scenario 1, when only punishment is possible, punishment-signalling evolves, regardless of the informativeness of small helping costs I_{SH}. **a**, Signallers are likely to punish when the punishment cost is small and **b**, Choosers are likely to accept Signallers who punish, while they almost always reject those who do not. In scenario 2, when only helping is possible, helping-signalling evolves, and becomes stronger as I_{SH} increases. **c**, Signallers are increasingly likely to help when the helping cost is small and **d**, Choosers are increasingly likely to accept Signallers who help, while they almost always reject those who do not. In scenario 3, when both signals are available, agents evolve to use both signals with equal frequency when they are equally informative, but to favour helping as I_{SH} increases. **e**, As I_{SH} increases, Signallers are increasingly likely to help, both when they have only a small helping cost (light blue dots), and when they have both small costs (dark blue dots); and are decreasingly likely to pay to punish, both when they only have a small punishing cost (light red dots), and when they have both small costs (dark red dots). **f**, As I_{SH} increases, Choosers are increasingly likely to accept Signallers who help but do not punish (blue dots), and increasingly likely to reject Signallers who punish but do not help (red dots). Furthermore, regardless of I_{SH}, Choosers almost always reject Signallers who neither help nor punish (brown dots). However, Chooser behaviour in response to Signallers who both punish and help (purple dots) stays at chance levels across all values of I_{SH} (because Signallers never send both signals, and thus Choosers do not face selection pressure to respond optimally to such Signallers).



Extended Data Figure 3 | Our two-stage experimental design involving Signallers and Choosers. First, in the signalling stage, the Signaller participates in a third-party punishment game (TPPG). Here a Helper decides whether to share with a Recipient, and then a third-party Punisher decides whether to pay to punish the Helper if the Helper was selfish (chose not to share). In our three experimental conditions, we manipulate the role(s) the Signaller plays in the TPPG. In the punishment-only condition, the Signaller plays once as the Punisher; in the punishment-plus-helping condition, the Signaller plays twice (with two different sets of other people) as the Punisher and the Helper; in the helping-only condition, the Signaller plays once as the Helper. Thus we vary which signal(s) are available. Second, in the partner choice stage, the Chooser plays a trust game with the Signaller. The Chooser decides how much to send the Signaller and any amount sent is tripled by the experimenter. The Signaller then decides how much of the tripled amount to return. Choosers use the strategy method to condition their sending on Signallers' TPPG decisions.



Extended Data Figure 4 | Third-party punishment is perceived as a stronger signal of trustworthiness than retaliation in our additional experiment (study 2). In our additional experiment, we manipulate whether the second stage of our game is a trust game (TG) or an ultimatum game (UG). In the TG, Choosers maximize their payoffs by sending more money to trustworthy Signallers (who will return a large amount); thus, preferential sending to punishers reflects expectations of punisher trustworthiness. In this game (left bars), punishment has large reputational benefits: replicating study 1, Choosers ($n = 405$) send 16 percentage points more to punishers than non-punishers, $P < 0.001$. In the UG, Choosers ($n = 421$) maximize their payoffs by sending more money to retaliatory Signallers (who are willing to pay the cost required to reject low offers); thus, preferential sending to punishers reflects expectations of punisher retaliation. In this game (right bars), punishment has smaller reputational benefits: Choosers send 3 percentage points more to punishers than non-punishers, $P = 0.001$. This difference between conditions is significant ($P < 0.001$) and robust to accounting for the fact that there is less overall variance in UG offers than TG transfers (see Supplementary Information section 6). Thus TPP is perceived as a stronger signal of trustworthiness (in the TG) than willingness to retaliate (in UG). These findings provide further evidence that our TG experiment results (study 1) are not driven by a perception that TPP signals retaliation (although TPP may also signal retaliation in other contexts). Shown is mean sending in each game. Error bars are ± 1 s.e.m.