nature human behaviour

LETTERS

PUBLISHED: 10 JULY 2017 | VOLUME: 1 | ARTICLE NUMBER: 0138

Corrupting cooperation and how anti-corruption strategies may backfire

Michael Muthukrishna[1,2*], Patrick Francois[3,4], Shayan Pourahmadi[5] and Joseph Henrich[2,3,4,6]

Understanding how humans sustain cooperation in large, anonymous societies remains a central question of both theoretical and practical importance. In the laboratory, experimental behavioural research using tools like public goods games suggests that cooperation can be sustained by institutional punishment—analogous to governments, police forces and other institutions that sanction free-riders on behalf of individuals in large societies[1–3]. In the real world, however, corruption can undermine the effectiveness of these institutions[4–8]. Levels of corruption correlate with institutional, economic and cultural factors, but the causal directions of these relationships are difficult to determine[5,6,8–10]. Here, we experimentally model corruption by introducing the possibility of bribery. We investigate the effect of structural factors (a leader's punitive power and economic potential), anti-corruption strategies (transparency and leader investment in the public good) and cultural background. The results reveal that (1) corruption possibilities cause a large (25%) decrease in public good provisioning, (2) empowering leaders decreases cooperative contributions (in direct opposition to typical institutional punishment results), (3) growing up in a more corrupt society predicts more acceptance of bribes and (4) anti-corruption strategies are effective under some conditions, but can further decrease public good provisioning when leaders are weak and the economic potential is poor. These results suggest that a more nuanced approach to corruption is needed and that proposed panaceas, such as transparency, may actually be harmful in some contexts.

Cooperation, particularly large-scale anonymous cooperation, remains an important puzzle to both evolutionary and social scientists, with real-world social and economic implications. One method for sustaining cooperation that has received considerable attention involves costly punishment[11–13], whereby individuals pay a cost to punish free-riders who fail to contribute to the public good. While cross-cultural evidence shows the ubiquity of costly punishment in large-scale societies (although not in small-scale societies), there is some variability in both the motivation to punish free-riders and the tendency to punish cooperators (for instance, some societies display significant levels of antisocial punishment—the punishment of cooperators)[14–16].

Research on the role of peer punishment in sustaining cooperation reveals two major challenges: (1) the second-order free-rider problem in which individuals defect on the job of punishing and thereby increase their payoffs[17,18] and (2) the problem of counter-punishment—punishment as revenge for previously being punished[12,19]. Institutional, or pool, punishment resolves these problems by designating one individual as a leader who can extract taxes and punish free-riders on behalf of other players[2]. Institutional punishment reduces the problems of both second-order free riding and counter-punishment, and may thus be important in explaining the emergence and maintenance of large-scale cooperation[3]. Moreover, recent empirical research shows that participants (at least participants from western, educated, industrialized, rich and democratic (WEIRD) nations[20]) prefer institutional punishment to peer punishment[1,21].

Institutional punishment, as typically modelled in public goods games (PGGs), serves to incentivize player choices when contributing to the public pool, and works by constraining leader choices to either punishing players or doing nothing. In the real world, however, channels such as bribery, nepotism and lobbying allow individuals (or corporations) to avoid contributing to the public pool (for example, by evading taxes) and to avoid being punished (for example, by paying a bribe instead). In other words, real-world leaders and institutions are corruptible.

Corruption is widespread, unevenly distributed and costly. The World Bank estimates that worldwide, US$1 trillion is paid in bribes alone[7]. However, the levels of corruption vary considerably. In Kenya, estimates suggest that 8 out of 10 interactions with public officials require a bribe and that the average urban Kenyan pays a bribe 16 times per month[22]. In contrast, the average Dane may never pay a bribe in their lifetime as Denmark has the lowest level of corruption based on the Corruption Perceptions Index[23]. The predicted costs of corruption vary from reductions in food redistribution anti-poverty programmes[24] to deaths from collapsed buildings[4]. Most recently, corruption has been identified as a contributing factor to the Greek economic crisis. Greece has the highest level of corruption in the European Union, with recent estimates placing its levels of corruption close to those of China and Brazil[23]. Corruption in European Union states, such as Greece, potentially undermines the future of the European Union. Although levels of corruption correlate with institutional, economic and cultural factors, the causal interconnections among these factors remain difficult to disentangle[8,9,25].

To model corruption, we modified the institutional punishment PGG (IPGG). In a PGG, players are given an endowment, which they can divide between themselves and a public pool. The public pool is multiplied by some amount and then divided equally among the players regardless of contribution. A cooperative dilemma is created by setting the multiplier such that it is in every player's best interest to allow others to contribute while contributing nothing themselves, but in the group's best interest for all players to contribute their entire endowment so that they all reap the maximum benefits of the multiplier. In the IPGG, one player is randomly selected

[1]Department of Psychological and Behavioural Science, London School of Economics and Political Science, London WC2A 2AE, UK. [2]Department of Human Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [3]Vancouver School of Economics, University of British Columbia, Vancouver, British Columbia V6T 1L4, Canada. [4]Canadian Institute for Advanced Research, Toronto, Ontario M5G 1M1, Canada. [5]Department of Economics, Columbia University, New York, New York 10027, USA. [6]Department of Psychology, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada. *e-mail: m.muthukrishna@lse.ac.uk

Table 1 | Leader decisions based on economic potential, leader strength and corruption exposure scores.

	Accept bribe	Punish	Do nothing
High economic potential	1.37 (0.65–2.21)	0.79 (0.41–1.14)	0.81 (0.29–1.40)
Strong leader	2.14 (1.18–3.36)	1.08 (0.60–1.61)	0.29 (0.10–0.50)
Player exposure corruption score	1.22 (1.01–1.44)	0.99 (0.81–1.19)	0.79 (0.63–1.02)
Player heritage corruption score	0.65 (0.54–0.79)	1.17 (0.96–1.40)	1.55 (1.25–1.89)
(Intercept)	0.57 (0.05–1.54)	0.16 (0.02–0.39)	3.01 (0.12–9.50)
Observations	1,396	1,396	1,396
n	175	175	175
Groups	45	45	45
Deviance information criterion	36.13	18.23	18.45

Values are reported as odds ratios and highest posterior density 95% confidence intervals. Odds ratios were estimated using a Markov chain Monte Carlo categorical generalized linear mixed model regression with the behaviour coded as 1 and the other two behaviours coded as 0. Each model regressed the behaviour in the BG (with no transparency or leader investment) on economic potential (low versus high), leadership strength (weak versus strong), and both player's and leader's exposure corruption score (*z* score) and heritage corruption score (*z* score), controlling for period, order of conditions, order of background questions, group size, age and gender with random effects for individuals within groups. Here, we report only the predictors of interest. The full model is reported in the Supplementary Information.

as a leader who can allocate punishments using taxes extracted from other players. Past research has shown the effectiveness of assigning designated leaders as institutional punishers[1,2,21].

To introduce bribery, we modified the IPGG by giving players and leaders one additional choice, thereby creating the bribery game (BG). In this scenario, in addition to dividing their endowment between themselves and the public pool, players can also offer some of their endowment to improve the leader's payoff (that is, effectively offering a bribe, although we use neutral language). In turn, leaders have an additional exclusive choice in addition to punishing or doing nothing to players: they can choose to take the contribution (that is, accept the bribe) or not. We chose to make punishing, accepting bribes or doing nothing to each player an exclusive choice for simplicity and because past research suggests that a non-exclusive choice would reduce or remove the impact of the bribe on decision-making[10]—in reality, a bribe with no effect would not last long. A new leader was selected in each round to remove any reputational effects, which turned the game into a series of repeated one-shot encounters. We manipulated the pool multiplier (a proxy for economic potential) and the punishment multiplier (the power of the leader to punish). In the BG, we also introduced three corruption mitigation strategies: partial transparency (revealing leader contributions), full transparency (revealing all leader behaviour, including bribe taking) and leader investment (forcing leaders to contribute their endowment to the public pool). We focus on transparency and discuss leader investment, which requires further investigation, in the Supplementary Information. We ran the experiment using a Canadian economic subject pool open to the public, which included native-born Canadians and first- and second-generation immigrants with diverse backgrounds.

We assumed players: (1) brought cultural differences to the game, which were shaped by their different ethnic backgrounds and cultural exposure; and (2) adjusted their behaviours via exposure to the experimental setting, moving closer to the equilibrium that maximized payoffs. We modelled an IPGG with a fixed tax rate to more realistically capture a world in which taxes were not directly correlated with punishment and where leaders could punish without a large cost to themselves (since their own taxes were a small part of the taxes contributing to the pool punishment or institution). We then modified the game to turn it into a BG by offering players and leaders the choice to offer and accept bribes. Without punishment, contributions tend towards zero. This is because contribution levels are contingent on the strength of leaders and their tendency to punish low contributors. We predicted that leaders would use taxes as punishment in the IPGG, since they are not personally costly and they benefit the leader's payoff by increasing the size of the public good. With increased leader strength, we predicted higher contributions and more public good provisioning. With regards to the BG, we predicted that players would have no incentive to offer contributions or bribes unless they were punished for not doing so. However, when bribery was an option, leaders would have a greater incentive to punish people for not offering brides than for not contributing, since their share of the public good would be smaller than a bribe multiplied by every player. More power gives leaders an increased ability to impose their will, increasing the rate of bribery at the expense of the public good. Thus, in contrast to the IPGG, we predicted that stronger leaders in the BG would reduce contributions and public good provisioning. However, if players had a preference for contributions over bribes (for example, if their previous experience was a world where potential returns on the public good were higher or where anti-corruption norms were adaptive), the incentive to punish bribes over contributions would be dampened. In contrast, growing up in a more corrupt society may lead to a higher preference for eliciting, offering and accepting bribes. Our full set of predictions is provided in the Supplementary Information.

To examine the costs of corruption, we compared the IPGG and BG. We found that when bribery was an option, mean contributions dropped by 25%. The difference between these conditions (estimated using a Markov chain Monte Carlo generalized linear mixed model regression; Supplementary Table 2) represented a 0.43 (95% confidence interval: −0.49 to −0.38) s.d. loss (1.4 points per period, equivalent to 14% of the initial endowment or Canadian $2.10 over the course of the game). Not surprisingly, when corruption could enter, it did, and cooperation deteriorated.

Having established the impact of bribery on cooperation, we examined the causes of this corruption. In Table 1 and Fig. 1 we used a Markov chain Monte Carlo categorical generalized linear mixed model regression to estimate the effect of (1) our different treatments, (2) cultural experience and (3) background on leader decisions. Leaders with a stronger punishment multiplier at their disposal (that is, stronger leaders) were about twice as likely to accept bribes and about three times less likely to do nothing. In contrast, when accepting bribes was not an option (that is, in the IPGG), the more powerful leaders were as likely to do nothing (see 'Leader decisions' in Supplementary Information). Thus, as expected, more power led to more corrupt behaviour.

Exploring individual variation, we found that those who grew up in more corrupt countries were more willing to accept bribes. For every one s.d. increase in players' exposure corruption scores (see 'Corruption perception scores' in Supplementary Information for

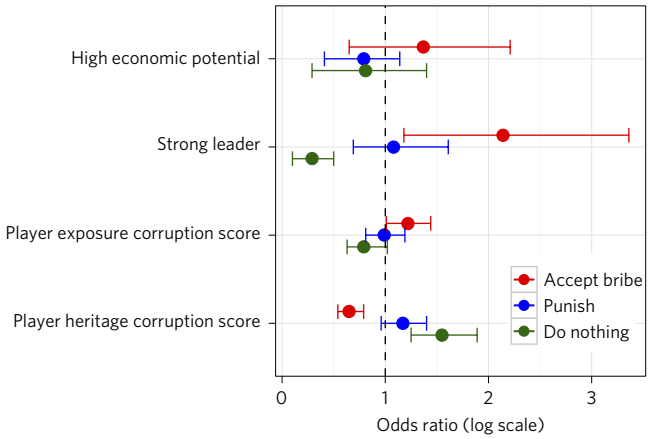

Figure 1 | Leader decisions based on economic potential, leader strength and corruption exposure scores. Odds ratios and 95% confidence intervals are shown for each behaviour (accept bribe, punish or do nothing).

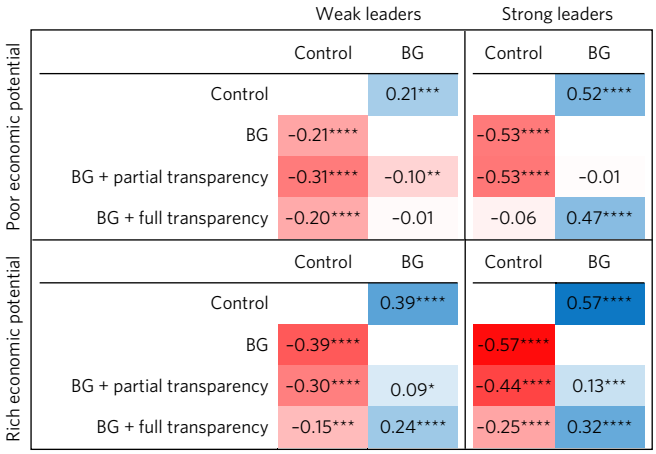

Figure 2 | Cures for corruption when there is a weak versus strong leader and when there is rich versus poor economic potential. Darker blue indicates greater public goods provisioning and darker red indicates reduced public goods provisioning. All coefficients were extracted from a single model by changing reference groups. The columns represent the reference group treatment (control versus BG), while each row shows the coefficient of each treatment compared with this reference group. The contributions were *z* scores, so the coefficients represent s.d. The full model is reported in the Supplementary Information. In all models, we accounted for the clustering inherent in the experimental design by including a fixed effect for the number of subjects and random effects for participants within groups. Note that in all treatments and structural contexts, the BG has lower contributions than the structurally equivalent IPGG (control). Corruption mitigation effectively increases contributions (although not to control levels) when leaders are strong or the economic potential is rich. When leaders are weak and the economic potential is poor, the apparent corruption mitigation strategy, full transparency has no effect and partial transparency further decreases contributions. $^{*}P < 0.10$; $^{**}P < 0.05$; $^{***}P < 0.01$; $^{****}P < 0.001$.

details on how these scores were constructed and the distribution of these scores in our sample), leaders were 1.2 times more likely to accept a bribe. In contrast, when players' parental heritage included countries with higher corruption norms (that is, more perceived corruption), leaders were 1.5 times less likely to accept bribes for every s.d. increase in corruption score and 1.6 times more likely to do nothing (see Fig. 1; the Supplementary Information shows all the models). In combination with other evidence[5,6,26–29], we suspect that our corruption exposure scores captured internalized social norms related to corruption acquired while growing up in different communities. Meanwhile, our parental heritage effects, which were driven by the Canadian-born participants (for example, second-generation immigrants), may have captured an internalized reaction against ethnic stereotyping—for instance, a reaction against the assumption of corrupt behaviour from those of their heritage[21].

Having generated corruption, we attempted to suppress it by modifying the BG using two different forms of transparency measures and by forcing leaders to invest in the public good. The first transparency approach, partial transparency, allowed all players to see the leader's contribution, thereby offering leaders an opportunity to establish or reveal a norm by revealing to players how much or how little leaders invested in the public pool. The second transparency approach, full transparency, allowed players to see all leader actions: leader contributions, the anonymized contributions and bribes from each player, and the leader's decision in each case. Leader investment forced leaders to maximally contribute their endowment to the public good, thereby tying a large part of their payoff to the efficiency of the public good. Tying leader payoffs to the success of the public good was explicitly used as one aspect of an anti-corruption measure in Singapore, which has one of the lowest levels of corruption (based on the Corruption Perceptions Index[23]) and the highest-paid leader in the world[30]. Singaporean minister salaries are pegged at the salaries of top professionals and Singapore's gross domestic product. The leader investment treatment was designed to be similar to linking leader payoffs to a country's gross domestic product, but in a way that minimally deviated from the other treatment designs. This treatment, though interesting, has certain caveats in its interpretation and requires further investigation. We report its effect and discuss these issues in more detail in the Supplementary Information.

To determine the effectiveness of these anti-corruption measures, we compared contributions in each condition to the IPGG (control) and BG. We regressed contributions (*z* scores) on treatment, economic potential and leader strength. The results of this regression are shown in Fig. 2 and separate coefficients within each condition can be seen. Note that these values come from a single model and were calculated by changing reference groups (see Supplementary Information). The raw mean contribution values are shown in Fig. 3.

Figures 2 and 3 reveal that stronger leaders were better able to increase the efficiency of public goods provisioning when the economic potential was poor and bribery was not an option (red bars in the top row), but when bribery was an option (blue bars) stronger leaders in poor contexts reduced the efficiency of the public good, making themselves wealthy at the expense of other players. Corruption mitigation effectively increased contributions (although not to control levels) when leaders were strong or the economic potential was rich. When leaders were weak and the economic potential was poor, the apparent corruption mitigation strategy, full transparency, had no effect and partial transparency further decreased contributions to levels lower than the standard BG (leading to less public good provisioning).

Although the cost of bribery was seen in all contexts, in poor economic contexts, the already low contributions were reduced even further. That is, even if powerful leaders were accepting bribes at comparable levels in both poor and rich economic contexts, the degree of corruption was not as visible if the economic potential was high. Leaders in richer economic contexts, such as the United States, may accept 'bribes' in the form of lobbying or campaign funding, which may indeed reduce the efficiency of the public good, but this cost is not as obvious since the economic potential is already much higher than in other nations. In contrast, in poorer economic contexts, such as the Democratic Republic of the Congo,

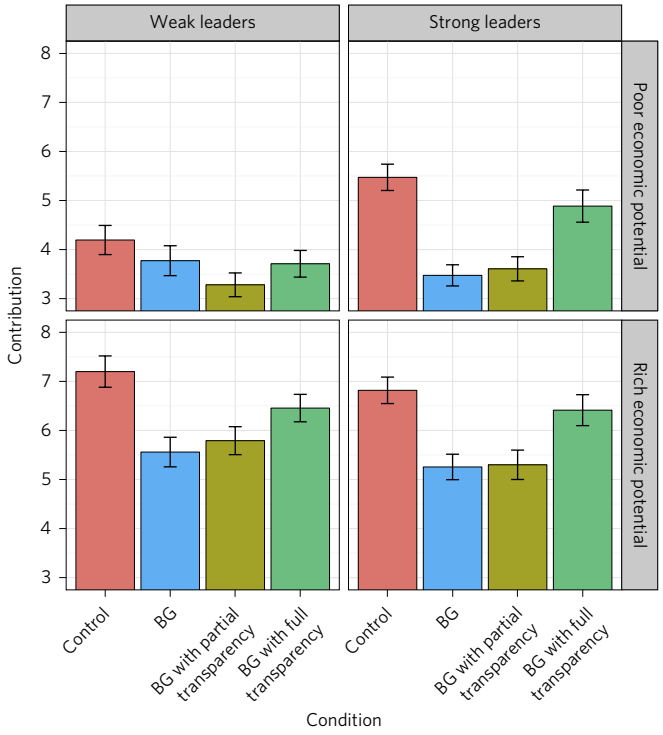

Figure 3 | Leader contributions by condition. Raw contributions (of the ten endowed points) and 95% confidence intervals for each within-subject treatment (control, BG, BG with partial transparency or BG with full transparency) in each between-subjects structural context (strong versus weak leader and poor versus rich economic potential). These data are consistent with our theory that predicts that more powerful leaders increase contributions in the IPGG but decrease contributions in the BG.

corruption further reduces the already low public good provisioning. Unfortunately, our results suggest that in these contexts with weak institutions and poor economic potential, efforts to mitigate corruption, such as transparency or leader investment, could backfire, further reducing investments in the public good. These results reflect leaders lacking the power to increase contributions through punishment and thus recouping the cost of their investment in the public good by accepting bribes. Transparency in this context reveals a low contribution norm. Thus, the lessons in fighting corruption when institutions have the power to sustain public goods (if only corruption were reduced) and the potential for economic growth is high may not only fail to apply when these conditions are not met, but could worsen the situation.

Our results suggest that the effect of exposure to different institutions and norms persists after moving to a new environment. This increase in corrupt behaviour following direct exposure to corrupt institutions or norms is consistent with the internalization of perceived norms[5,6,26,27] and with previous empirical data showing, for example, that diplomats from high-corruption countries accumulate more unpaid parking violations[29]. However, the decreased probability of accepting bribes among those whose cultural background includes more-corrupt countries suggests that second-generation and later migrants are not as corrupt as their peers from less-corrupt nations. This may represent the self-selection of immigrants from their home countries or may be a form of 'identity denial'[21], whereby acculturated individuals actively avoid the stereotypes of their inherited ethnic labels. Although we used a large range of corruption scores (see 'Corruption perception scores in Supplementary Information), our sample was limited to migrants in a Canadian context and further investigation is required to determine if these cultural results can be generalized. Together, these results suggest that corruption may be rooted in structural factors, but that internalized corruption norms may cause these behaviours to persist in a new context.

Overall, these results suggest that: (1) stronger institutions and leaders are required to sustain public goods contributions when the economic potential is poor and the incentive to free ride is high; (2) in this context, when they are able to, leaders abuse their power with a noticeable economic cost; and (3) despite this, even if the economic potential is poor, if leaders are powerful, anti-corruption measures can be effective at increasing public good provisioning. Thus, efforts to mitigate corruption in poorer economic contexts must go hand in hand with strengthening institutions. When leaders have less punitive power, efforts such as transparency may have no effect or even decrease contributions as they reveal the rationality of low public good contributions and show that most leaders do not contribute. In a rich context with powerful punitive institutions, there may be multiple equilibria that just require norms (activated in our game by transparency) to stabilize a higher payoff. In contrast, in a poor context with weak institutions, there is only one equilibrium: bribe offers and low public good provisioning.

Although these experimental results begin to offer insights into the causal effect of corruption on cooperation, extending such experimental findings demands great caution. Laboratory work on the causes and cures of corruption must inform and be informed by real-world investigations of corruption from around the globe. Thus, aiming only to drive future investigations, our results suggest that as the economic potential grows, less government intervention is required to enforce cooperation and increased power may be misused, requiring greater anti-corruption efforts. In contrast, when the economic potential is poor, strong government intervention is most effective at decreasing free riding, as long as this intervention is paired with strategies to mitigate corruption. This may help explain why intuitions about 'cures for corruption' based on experiences in rich nations do not work as well in poorer nations.

Methods

Participants. A total of 274 participants (166 females; mean age: 20.90), drawn from an economic subject pool open to the public, took part in the study. Their ethnic backgrounds were as follows: 63 European Canadians, 158 East Asians, 17 South Asians and 36 of other ethnicities. The participants played in groups of between four and seven players. Ethical approval was obtained from the University of British Columbia Behavioural Research Ethics Board (H12-02457). Informed consent was obtained from all participants before the start of the study. The participants were randomly assigned to the experimental groups.

Experimental design. We used a 2 (high versus low economic potential) × 2 (weak versus strong leader power) between-subjects experimental design with five within-subject treatments: IPGG control ($n = 205$), BG ($n = 222$), BG with partial transparency ($n = 228$), BG with full transparency ($n = 204$) and BG with leader investment ($n = 196$). Allocation to all treatments was random. The sample sizes for the four between-subjects treatments were as follows: low economic potential and weak leader power ($n = 71$), low economic potential and strong leader power ($n = 68$), high economic potential and weak leader power ($n = 68$) and high economic potential and strong leader power ($n = 67$).

In the real world, leaders make institutional decisions based on a fixed budget to which they are one among many contributors and which has to be spent. To better model these conditions, we extracted fixed taxes for punishment, which were discarded if not used. Participants were randomly assigned to one of the four between-subjects treatments and four of the five within-subjects treatments.

To test the possible contributing causes of corruption, we randomly assigned each group of participants to a treatment with (1) either a high or low marginal per capita rate of return (0.3 versus 0.6) as a measure of economic potential and (2) either a high or low punishment multiplier (1 versus 3) as a measure of the strength of the leader or institution. The marginal per capita rate of return was the expected return for every point invested in the public pool and the punishment multiplier was the number of points subtracted from a sanctioned player for every tax point spent on punishing that player.

The within-subject treatments were played in a random order with pre-recorded video instructions before each period. A quiz was conducted at the start to ensure participants knew how each treatment worked. This quiz, along with the script and screenshots from the video, is in the Supplementary Information.

We used a block randomization design, in which participants played a minimum of ten rounds, but the game could end at any point before the completion of ten rounds. At ten rounds, the participants were informed which round the period had ended at or played further rounds until the game ended. In this way, there were ten rounds to analyse without end-game effects—that is, participants did not know when the game would end. To remove reputational effects, the leader was also randomly selected for each round. Replacement was performed by random selection, such that players also could not say that the same person could not be the leader for a consecutive round. As such, the experiment could be interpreted as a series of one-shot interactions. The participants were paid for ten random rounds from across all the conditions. They were paid at a rate of 15c per point, with a show up fee of $10.

Measures. We measured age, gender, university degree or occupation and major or industry, prestige/dominance, right wing authoritarianism, whether participants had spent their entire life in Canada, where else they had lived, which suburb they had grown up in, ethnic group, religion and importance of religion, how well they spoke the language of their ethnic heritage (cultural competence), inclusion of other in the self scale (identification with their ethnic group and identification with Canadians), the Vancouver Index of Acculturation, and mainstream versus heritage acculturation (integration into culture).
Two corruption scores were calculated for each participant using the mean of Transparency International's Corruption Perceptions Index for all the countries each participant had lived in and all the countries from which they derived their ethnic heritage. The Corruption Perceptions Index has a scale from 0 (most corrupt) to 100 (least corrupt). For each country, we subtracted this value from 100 (so that higher scores indicated higher corruption). Perception of corruption was chosen as the measure of corruption as it indicated the perceived norm for national corruption.

The heritage corruption score primarily represents the potential influence of vertically transmitted corruption norms (parent to child), whereas the exposure corruption score represents corruption norms that the participant may have acquired through non-parental cultural transmission or direct experience.

We asked the last 39 groups (194 participants) their preferences for the conditions of the game. These participants were asked these questions after all other measures had been taken so that there were no differences in experimental design between them and the preceding 17 groups (79 participants). We report these preferences, along with the details of all the measures in the Supplementary Information.

Data availability. The data that support the findings of this study are available in figshare with the identifier https://doi.org/10.6084/m9.figshare.5004956.

Received 13 July 2016; accepted 7 June 2017; published 10 July 2017

References

1. Fehr, E. & Williams, T. *Endogenous Emergence of Institutions to Sustain Cooperation* (2013); http://researchers-sbe.unimaas.nl/wp-content/uploads/gsbe/spring-2014/papers-and-abstracts/Paper_williams.pdf
2. O'Gorman, R., Henrich, J. & Van Vugt, M. Constraining free riding in public goods games: designated solitary punishers can sustain human cooperation. *Proc. R. Soc. Lond. B Biol. Sci.* **276,** 323–329 (2009).
3. Sigmund, K., De Silva, H., Traulsen, A. & Hauert, C. Social learning promotes institutions for governing the commons. *Nature* **466,** 861–863 (2010).
4. Ambraseys, N. & Bilham, R. Corruption kills. *Nature* **469,** 153–155 (2011).
5. Gächter, S. & Schulz, J. F. Intrinsic honesty and the prevalence of rule violations across societies. *Nature* **531,** 496–499 (2016).
6. Weisel, O. & Shalvi, S. The collaborative roots of corruption. *Proc. Natl Acad. Sci. USA* **112,** 10651–10656 (2015).
7. Kaufmann, D. Myths and realities of governance and corruption. http://dx.doi.org/10.2139/ssrn.829244 (2005).
8. Treisman, D. What have we learned about the causes of corruption from ten years of cross-national empirical research? *Annu. Rev. Polit. Sci.* **10,** 211–244 (2007).
9. Treisman, D. The causes of corruption: a cross-national study. *J. Pub. Econ.* **76,** 399–457 (2000).
10. Gneezy, U., Saccardo, S. & van Veldhuizen, R. Bribery: greed versus reciprocity. http://dx.doi.org/10.2139/ssrn.2803623 (2016).
11. Fehr, E. & Gachter, S. Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* **90,** 980–994 (2000).
12. Gächter, S., Renner, E. & Sefton, M. The long-run benefits of punishment. *Science* **322,** 1510–1510 (2008).
13. Rockenbach, B. & Milinski, M. The efficient interaction of indirect reciprocity and costly punishment. *Nature* **444,** 718–723 (2006).
14. Henrich, J. *et al.* Costly punishment across human societies. *Science* **312,** 1767–1770 (2006).
15. Henrich, J. *et al.* Markets, religion, community size, and the evolution of fairness and punishment. *Science* **327,** 1480–1484 (2010).
16. Herrmann, B., Thöni, C. & Gächter, S. Antisocial punishment across societies. *Science* **319,** 1362–1367 (2008).
17. Boyd, R. & Richerson, P. J. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol. Sociobiol.* **13,** 171–195 (1992).
18. Panchanathan, K. & Boyd, R. Indirect reciprocity can stabilize cooperation without the second-order free-rider problem. *Nature* **432,** 499–502 (2004).
19. Nikiforakis, N. Punishment and counter-punishment in public good games: can we really govern ourselves? *J. Pub. Econ.* **92,** 91–112 (2008).
20. Henrich, J., Heine, S. J. & Norenzayan, A. Most people are not WEIRD. *Nature* **466,** 29–29 (2010).
21. Traulsen, A., Röhl, T. & Milinski, M. An economic experiment reveals that humans prefer pool punishment to maintain the commons. *Proc. R. Soc. B Biol. Sci.* **279,** 3716–3721 (2012).
22. Kenya, T. I. *The Kenya Urban Bribery Index* (Transparency International—Kenya, 2001).
23. Transparency International. Corruption perceptions index 2014 brochure. https://www.transparency.org/cpi2014/results#myAnchor2 (2014).
24. Olken, B. A. Corruption and the costs of redistribution: micro evidence from Indonesia. *J. Pub. Econ.* **90,** 853–870 (2006).
25. Pande, R. & Olken, B. Corruption in developing countries. *Annu. Rev. Econom.* **4,** 479–509 (2012).
26. Chudek, M. & Henrich, J. Culture–gene coevolution, norm-psychology and the emergence of human prosociality. *Trends Cogn. Sci.* **15,** 218–226 (2011).
27. Chudek, M., Muthukrishna, M. & Henrich, J. in *The Handbook of Evolutionary Psychology* Vol. 2 (ed. Buss, D. M.) Ch. 30 (John Wiley and Sons, 2015).
28. Kimbrough, E. O. & Vostroknutov, A. *Norms Make Preferences Social* (Dept of Economics, Simon Fraser Univ., 2013).
29. Fisman, R. & Miguel, E. Corruption, norms, and legal enforcement: evidence from diplomatic parking tickets. *J. Polit. Econ.* **115,** 1020–1048 (2007).
30. Kuan Yew, L. *From Third World to First World; the Singapore Story: 1965–2000* (Harper Collins, 2000).

Acknowledgements

J.H. acknowledges support from the Canadian Institute for Advanced Research. J.H. and P.F. acknowledge support from the Social Sciences and Humanities Research Council, Canada. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

M.M., P.F., S.P. and J.H. developed the theory, designed the experiments and wrote the paper. M.M. and S.P. carried out the experiments. M.M., S.P. and J.H. conducted the statistical analyses.

Additional information

Supplementary information is available for this paper.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to M.M.

How to cite this article: Muthukrishna, M., Francois, P., Pourahmadi, S. & Henrich, J. Corrupting cooperation and how anti-corruption strategies may backfire. *Nat. Hum. Behav.* **1,** 0138 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Competing interests

The authors declare no competing interests.