Indirect punishment and generosity towards strangers

Ule, A.; Schram, A.; Riedl, A.; Cason, T.N.

DOI
10.1126/science.1178883

Publication date
2009

Document Version
Author accepted manuscript

Published in
Science

Indirect Punishment and Generosity towards Strangers*

Aljaž Ule[1,2]**, Arthur Schram[1], Arno Riedl[3], Timothy N. Cason[4]

[1] Center for Research in Experimental Economics and Political Decision-making, University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, the Netherlands.

[2] Faculty of Mathematics, Natural Sciences and Information Technologies, University of Primorska, Glagoljaška 8, SI-6000 Koper, Slovenia.

[3] Department of Economics (AE1), Maastricht University, P.O. Box 616, 6200 MD Maastricht, the Netherlands.

[4] Department of Economics, Krannert School of Management, Purdue University, 100 S. Grant Street, West Lafayette, IN 47907-2076, USA.

All authors contributed equally to this work.

* *This is the final peer-reviewed manuscript version of the following article:*

Indirect Punishment and Generosity Toward Strangers

Aljaž Ule, Arthur Schram, Arno Riedl, Timothy N. Cason

***Science** (18 December) 2009: Vol. 326. no. 5960, pp. 1701 - 1704*

which has been published in final form at DOI: 10.1126/science.1178883

** To whom correspondence should be addressed. E-mail: a.ule@uva.nl

October 20, 2009

One-sentence summary:

We experimentally show that being generous towards strangers pays off, but only if unhelpful strangers can be punished.

Abstract:

Many people incur costs to reward strangers who have been kind to others. Theoretical and experimental evidence suggests that such 'indirect rewarding' sustains cooperation between unrelated humans. Its emergence is surprising, since rewarders incur costs but receive no immediate benefits. It can prevail in the long run only if rewarders earn higher payoffs than 'defectors' who ignore strangers' kindness. We provide experimental evidence regarding the payoffs received by individuals who employ these and other strategies, such as 'indirect punishment' that imposes costs on unkind strangers. We find that if unkind strangers cannot be punished defection earns most. If they can be punished, however, then indirect rewarding earns most. Indirect punishment plays this important role even if it gives a low payoff and is rarely implemented.

Indirect reciprocity is widespread in human societies. It occurs when we incur costs to reward those who we know have been kind to others or punish those who we know have been unkind to others. Indirect reciprocity is based on reputation and helps to enforce trustworthy behavior between individuals who do not know each other and who may not meet again. Such encounters form a substantial part of our interactions and are especially frequent in online commerce.

Indirect reciprocity is at work, for example, when someone financially supports anonymous volunteers working at food banks that help the poor, even though she does not face the risk of ever needing a food bank's services. The donors' helping behavior is therefore called indirect rewarding. Such costly indirect rewarding is thought to be a key factor in the evolution of human cooperation (*1-4*). Experimental research (*5-8*) and theoretical considerations (*9-11*) suggest that indirect rewarding can indeed sustain cooperation among unrelated humans. There is little empirical evidence about the long term performance of indirect rewarding itself, however. People who engage in costly rewarding may in the long run lose out against defectors who never reward and thus avoid the associated costs. Even if good reputation is rewarded (*12*) indirect rewarders might lose out against 'cautious defectors' who appear generous only to avoid a bad reputation.

Indirect reciprocity may also take the form of costly indirect punishment. Though punishment has been observed to be important for promoting cooperative behavior in direct encounters (*13*), recent theoretical work suggests that it may be only marginally relevant

when interaction is indirect (*14*). Empirical evidence on the use of indirect punishment and its long term performance is missing, however.

We provide experimental evidence of human behavior in an anonymous environment where individuals can indirectly reward and punish. We determine the occurrences of different types of behaviors, including indirect rewarding, indirect punishment, defection and cautious defection, among human subjects and determine their payoff performance.

Our experimental design builds on the so-called indirect helping game (*5,8,9*). In total, 140 participants are repeatedly (100 rounds), anonymously and randomly matched into donor-recipient pairs. Because roles are determined randomly, participants will typically be donor in approximately half of the rounds. In the indirect helping game, only donors make decisions. In any round, each donor first observes the recipient's recent behavior in the role of donor and then decides whether to 'help' the recipient, or to 'pass'. Helping is costly for the donor and beneficial for the recipient, with the benefits exceeding the costs. In earlier experiments, indirect punishment was not available as an option; a restriction that is arguably not a realistic feature of human interactions (*13-16*). In our experiment the donor can choose to 'hurt' the recipient instead of passing or helping. Hurting is costly for the donor, but we vary its impact on the receiver. We conducted two treatments that differ only in this impact, which allows us to isolate the effect of indirect punishment on the payoff performance of different types of behavior. In our main treatment (HP) a hurt recipient loses 250 units of our experimental money, 'francs'. In the control treatment (SP) a hurt recipient loses or earns no francs. We say that punishment is harmful in HP but only

symbolic in SP. In both treatments, the donor loses 50 francs for hurting or 200 francs for helping, and the recipient earns 250 francs when she receives help. Passing does not affect either player's payoff. In both treatments the recipient observes the donor's action. Treatment SP is a control for HP because it identifies differences in behavior between environments where indirect punishment has material consequences for the recipient and where it does not, while holding all other parameters constant across treatments (*17-19*).

Before choosing an action, donors observe a part of their recipient's donating history. A donor always learns her recipient's three most recent actions (1^{st}-order information) and can access for a small price the 1^{st}-order information their recipients observed when making these decisions (2^{nd}-order information). For treatment HP and SP we collected data for, respectively, 8 and 6 independent cohorts of 10 subjects.

The aggregate frequency (60.7%) of helping choices in our experiment falls within the range (50%–85%) observed in experiments without the option of indirect punishment (*5,8*). Comparing HP to SP surprisingly shows that, in spite of the possibility of imposing costs on uncooperative recipients in HP, the two treatments exhibit no significant differences in average helping rates. Donors choose help with 60.0% frequency on average across the 6 cohorts in SP, and 61.2% frequency on average across the 8 cohorts in HP. This difference is not significant (z=–0.065, P=0.95, 2-sided Mann-Whitney U-test, N=14). Since behavior in both treatments displays a pronounced endgame effect we restrict our analysis to the first 90 rounds (*17*).

In SP punishment is very rare (1.1%), which is not surprising because it is costly for the donor but only symbolic to the recipient. When punishment is harmful (HP) it is used significantly more often (z=−2.207, P=0.027, 2-sided Mann-Whitney U-test, N=14) but still infrequently (3.4%). In both treatments donors typically reward kind behavior with helping. When the recipient's history reveals unkind behavior towards others, donors more often pass than hurt. This preference for passing may explain why the punishment option in HP fails to increase cooperation beyond levels obtained without an option to punish (*5,8*). The infrequent use of punishment in our indirect reciprocity game seems to contrast the experimental results from public goods games with direct punishment, where frequent punishment of defectors sustains cooperation in the short *(13,20)* and intermediate run *(21)*. This difference in results might be driven by the structural differences between the games. In our indirect reciprocity game each action is indirect and targeted at a single person. By contrast, only punishment can be targeted at a specific person in public goods games, while each other action affects every member of the group. In combination with the difference in parameters, this may explain the level of punishment we observe *(17)*.

Recent evidence suggests that human reciprocity is driven to a large extent by stable behavioral strategies (*22-24*) and a rich set of such strategies has been identified in recent models of evolution of indirect reciprocity (*1,9,10*). We consider seven prominent behavioral strategies and assess their payoff performance. These strategies are partitioned along the different ways a donor may use her own history or that of the recipient when choosing her action (*10*).

The first, partition distinguishes between ‘discriminate’ and ‘indiscriminate’ strategies. An indiscriminate strategy does not condition an action on the donor’s or the recipient’s histories. For example, ‘indiscriminate altruism’ always prescribes help and ‘indiscriminate defection’ always prescribes pass.

The second partition divides discriminate strategies into those with selfish concern (‘self-regarding’) and those with concern for others (‘other-regarding’). The strategy ‘cautious defection’ employs occasional helping in order to maintain the donor's good reputation. In particular, it prescribes to help only when the donor's own history shows little helping. It is therefore discriminating and self-regarding (*10*).

Strategies that do condition actions on the recipient's history are discriminate and other-regarding. We focus on the ‘reciprocal’ strategies that prescribe help only to those recipients whose history reveals frequent helping. The third partition divides the reciprocal strategies between ‘punishing’ strategies that use the possibility to ‘hurt’ unhelpful recipients and the ‘rewarding’ strategies that do not.

The fourth partition divides the reciprocal strategies on the basis of the type and amount of information they use. This allows us to distinguish between standing and image scoring (*1,9,10,25*). An individual’s ‘image score’ and ‘standing’ are statistics that summarize her reputation (*9,10*). Her image score decreases when she passes and increases when she helps, while her standing decreases only when she passes on a recipient with a good reputation. An image scoring strategy prescribes helping only those recipients with a high image score (*9*) and a standing strategy prescribes helping only those with high standing

(*10*). Specifically, a standing strategy prescribes that a donor bases her action not only on the 1st-order information about the recipient but also on the underlying 2nd-order information. The latter indicates what the recipient knew about her recipient when choosing her past actions as a donor. Using the combination of the final two partitions we distinguish between 'image rewarding', 'image punishing', 'standing rewarding', and 'standing punishing'. Because of the availability of hurting and the limits on the observable history in our experiment we consider approximated standing and image scoring strategies.

For each participant, we determine whether her actions across rounds 1-90 are consistent with any single behavioral strategy. We give the details and a graphic depiction of our classification procedure in the supporting online material (*17*). We refer to classified subjects as rewarder, punisher, etc. Table 1 summarizes the identified strategies and shows for each the proportion of subjects using it. Almost all participants (SP: 86.7%, HP: 95.0%) can be classified. More classified subjects use image scoring strategies than standing strategies. Among them more are rewarders than punishers. The next largest fractions are those of indiscriminate and cautious defectors, with approximately half of them being cautious. The smallest group is formed by indiscriminate altruists.

Little is known about the payoff consequences of using various strategies in indirect reciprocity games (*12,26*). Such information is important because a strategy can flourish in the long run only if it yields a higher benefit than the alternatives. We consider the identified strategies and calculate the average cohort payoff each generates (Fig. 1). For each subject we calculate, in francs, her average earnings as a donor plus her average

earnings as a recipient across rounds 1-90. The payoff for a strategy is calculated for each cohort where the strategy is observed, as the average payoff across the subjects using it. These payoffs per cohort are used in our statistical analysis. It is the relative fitness of a strategy which determines its long-term success, however. Fig. 1 therefore shows for each treatment the average payoff of each strategy relative to the treatment average payoff. This relative payoff is calculated as ((average payoff of all subjects in treatment using a particular strategy)–(average treatment payoff))/(average treatment payoff).

Fig. 1 reveals important payoff differences between the two treatments. In treatment SP the indiscriminate defectors fare best (average payoff=23.69). Compared to the combined classes of defectors (20.96), the combined rewarders (13.28) earn significantly less (P=0.044, 2-sided Wilcoxon signed-ranks test, N=5 paired observations). The altruists fare worst (4.79). Hence, defection outperforms all other strategies. In treatment HP the payoffs are strikingly different. The cautious defectors (14.23), the image rewarders (14.80) and the standing rewarders (13.80) are more successful than the indiscriminate defectors (6.16). Even if we combine the two classes of defectors (10.20), the combined rewarders (14.50) earn significantly more (P=0.068, 2-sided Wilcoxon signed-ranks test, N=8, paired observations). Noticeably, the punishment strategies, which are used only in HP, are among the least successful (6.4).

The stark difference in the ranking of earnings across the two treatments is caused mainly by the distinctly lower earnings of indiscriminate defectors in HP, as compared to SP (z=1.715, P=0.086, 2-sided Mann-Whitney U-test, N=9). This is a direct consequence of

harmful punishment. The slightly higher punishment rate in HP (3.4%) than in SP (1.1%) is almost entirely directed towards defectors (SP: 1.6%, HP: 12.8%) (z=1.976, P=0.048, 2-sided Mann-Whitney U-test, N=9) and cautious defectors (SP: 1.2%, HP: 5.2%) (z=1.375, P=0.169, 2-sided Mann-Whitney U-test, N=8). Hence, even though harmful punishment is rare, it substantially reduces defectors' earnings and changes the ranking of earnings among strategies.

Our results regarding the effects of indirect punishment complement recent experimental research showing that costly direct punishment may disfavor individuals and groups in repeated direct interactions with strangers, at least in the short run (*27-29*). However, our earnings comparisons across treatments reveal that in indirect reciprocity games punishment does not need to be frequent to promote the relative success of reward strategies. Theoretical models of indirect punishment investigating its long term effects on cooperation are just starting to emerge (*14*). Our study can aid the development of such models by showing that indirect punishment, even if it is rare, can support human cooperation.

References and Notes

1. R. Sugden, *The Economics of Rights, Cooperation and Welfare.* (Blackwell, Oxford, 1986).
2. R.D. Alexander, *The Biology of Moral Systems.* (Aldine de Gruyter, New York, 1987).
3. M.A. Nowak, *Science* **314**, 1560 (2006).
4. E. Fehr, U. Fischbacher, *Nature* **425**, 785 (2003).
5. C. Wedekind, M. Milinski, *Science* **288**, 850 (2000).
6. M. Milinski, D. Semmann, H.-J. Krambeck, *Nature* **415**, 424 (2002).
7. D.Semmann, H.-J. Krambeck, M. Milinski, *Behav. Ecol. Sociobiol.* **56**, 248 (2004).
8. I. Seinen, A.J.H.C. Schram, *Eur. Econ. Rev.* **50**, 581 (2006).
9. M.A. Nowak, K. Sigmund, *Nature* **393**, 573 (1998).
10. O. Leimar, P. Hammerstein, *Proc. R. Soc. Lond. B* **268**, 745 (2001).
11. K. Panchanathan, R. Boyd, *Nature* **432**, 499 (2004).
12. C. Wedekind, V.A. Braithwaite, *Curr. Biol.* **12**, 1012 (2002).
13. E. Fehr, S. Gächter, *Nature* **415**, 137 (2002).
14. H. Ohtsuki, Y. Iwasa, M.A. Nowak, *Nature* **457**, 79 (2009).
15. S. Bowles, H. Gintis, *Theor. Popul. Biol.* **65**, 17 (2004).
16. E. Fehr, U. Fischbacher, *Evol. Hum. Behav.* **25**, 63 (2004).
17. Materials and methods are available as supporting material on *Science* Online.
18. J. Carpenter, A. Daniere, L. Takahashi, *J. Econ. Org. Beh.* **55**, 533 (2004).

19. D. Masclet, C. Noussair, S. Tucker, M.C. Villeval, *Am. Econ. Rev.* **93**, 366 (2003)

20. B. Rockenbach, M. Milinski, *Nature* **444**, 718 (2006).

21. S. Gächter, E. Renner, M. Sefton, ***Science*** **322**, 1510 (2008).

22. R. Kurzban, D. Houser, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 1803 (2005).

23. E.H. Hagen, P. Hammerstein, *Theor. Popul. Biol*. **69**, 339 (2006).

24. B. Wallace, D. Cesarini, P. Lichtenstein, M. Johannesson, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 15631 (2007).

25. M. Milinski, D. Semmann, T.C.M. Bakker, H.-J. Krambeck, *Proc. R. Soc. Lond. B* **268**, 2495 (2001).

26. M.A. Nowak, K. Sigmund, *Nature* **437**, 1291 (2005).

27. A. Dreber, D.G. Rand, D. Fudenberg, M.A. Nowak, *Nature* **452**, 348 (2008).

28. M. Egas, A. Riedl, *Proc. R. Soc. Lond. B* **275**, 871 (2008).

29. B. Herrmann, C. Thöni, S. Gächter, ***Science*** **319**, 1362 (2008).

30. We are grateful for financial support by the Dutch science foundation NWO (Veni 451-07-031 and Evolution & Behavior 051-12-012) and would like to thank M. Egas, E. Fehr, S. Gächter, M. Milinski, M. Sabelis, J. Ule and M. van Veelen for helpful comments.

Supporting Online Material

www.sciencemag.org

Materials and Methods

Table 1. Strategies in indirect reciprocity game with punishment. The percentages of individuals identified with a strategy are given in parentheses, with their percentage in SP shown first and their percentage in HP shown second.

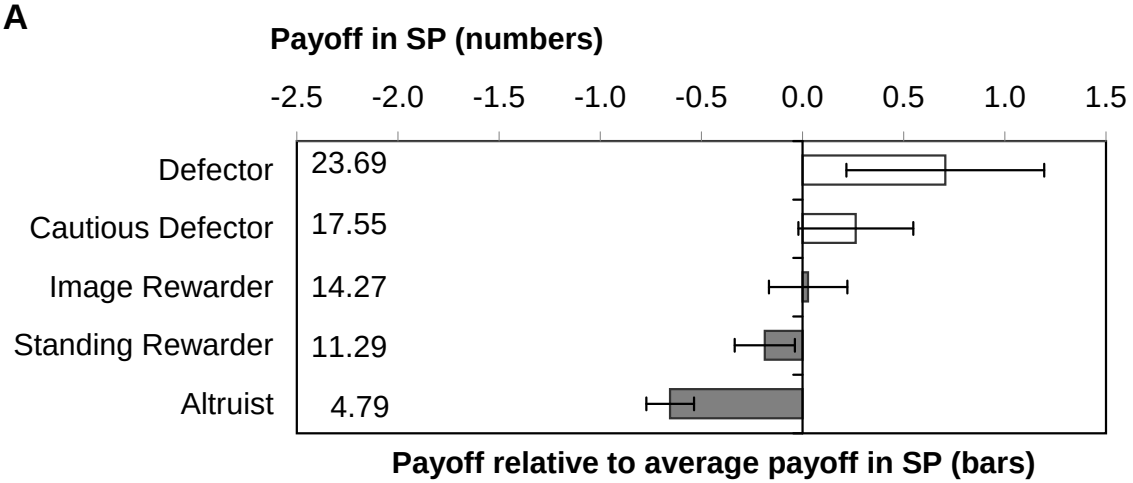
	self-regarding	*other-regarding*			
indiscriminate	defectors (9.6 , 10.5)	altruists (7.7 , 9.2)			
discriminate	cautious defectors (7.7 , 10.5)	rewarders		punishers	
		image (50.0 , 39.5)	standing (25.0 , 17.1)	image (0 , 5.3)	standing (0 , 7.9)

Figure Caption

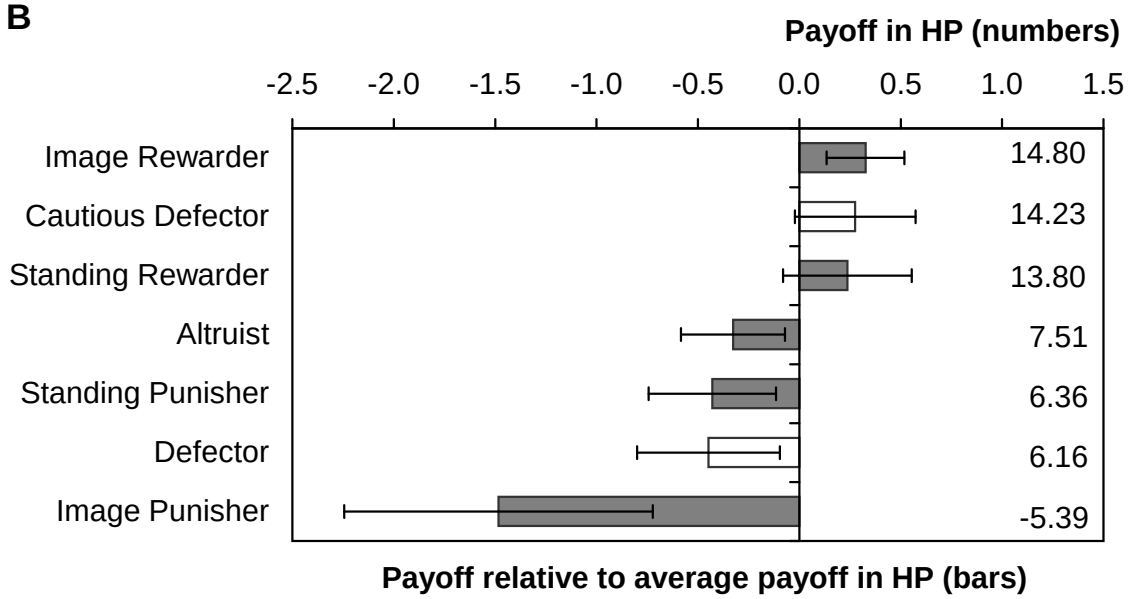
(figure file 1178883fig1.eps)

Fig. 1: Average and Relative Payoffs of the Different Strategies in (**A**) SP and in (**B**) HP. Numbers indicate average payoffs per round in francs of different strategies. Bars indicate the average payoffs of different strategies relative to the average payoffs across all subjects in the respective treatment, and error bars indicate +/- one standard error of these average relative payoffs. Tables S1-S3 in the Supplementary Online Material (*17*) provide detailed information of payoffs for each strategy, cohort and individual.

A


Payoff in SP (numbers)
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Defector
Cautious Defector
Image Rewarder
Standing Rewarder
Altruist
23.69
17.55
14.27
11.29
4.79
Payoff relative to average payoff in SP (bars)

B


Payoff in HP (numbers)
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Image Rewarder
Cautious Defector
Standing Rewarder
Altruist
Standing Punisher
Defector
Image Punisher
14.80
14.23
13.80
7.51
6.36
6.16
-5.39
Payoff relative to average payoff in HP (bars)