

HUMAN COOPERATION

Second-order free-riding problem solved?

Arising from: K. Panchanathan & R. Boyd *Nature* 432, 499–502 (2004)

Panchanathan and Boyd¹ describe a model of indirect reciprocity in which mutual aid among cooperators can promote large-scale human cooperation without succumbing to a second-order free-riding problem² (whereby individuals receive but do not give aid). However, the model does not include second-order free riders as one of the possible behavioural types. Here I present a simplified version of their model to demonstrate how cooperation unravels if second-round defectors enter the population, and this shows that the free-riding problem remains unsolved.

Suppose a population shows two types of behaviour. In each period, 'cooperators' pay a cost C to contribute to a public good and receive a benefit B . 'Defectors' also receive benefit B but do not pay the cost. Assuming growth of each type in the population is proportional to mean pay-offs, how can we explain the emergence and persistence of cooperators?

Panchanathan and Boyd propose that a first-round public-goods game is followed by a second-round 'mutual-aid' game. In this game, each 'shunner' cooperates in the first round and then pays a cost c to generate a benefit b to one randomly chosen shunner, which yields a pay-off of $B - C + b - c$. Shunners can invade and dominate a population of defectors if $b - c > C$. In other words, shunners will prevail if they can create an in-group net surplus from mutual aid in the second round that exceeds the cost of cooperation in the first round. However, this is only possible if shunners can exclude second-round defectors from mutual aid. Although Panchanathan and Boyd include in their model individuals who receive a benefit without paying the cost in the first round (defectors), they do not consider such a type for the second round. Thus, rather than solving the second-order free-rider problem, their model merely assumes it away.

To see why, suppose that 'second-round defectors' enter the population. These individuals receive the same pay-off as shunners in the first round and are eligible to receive aid in the second round. However, they do not give aid to others. If we denote the proportion of shunners in the population by p , then the average pay-off for second-round defectors is $B - C + bp$ and the average pay-off to shunners changes to $B - C + bp - c$. If second-round defectors invade a population of shunners, then second-round cooperation collapses, pay-offs to second-round defectors fall, and eventually ordinary defectors can invade and dominate the population — leaving us back where we started.

Panchanathan and Boyd implicitly acknowledge this problem when they note that each individual must have perfect information about the cooperation and aid-giving histories of all other members of the population in order for mutual aid to be sustained. In other words, shunners must be able to recognize and exclude second-round defectors from receiving aid. They do incorporate an error term into their model, but they do not consider errors in which individuals mistakenly help a recipient of bad reputation during the mutual-aid game¹.

Suppose that e is the probability that shunners mistakenly aid a second-round defector or withhold aid from a shunner. Cooperation can only be maintained when the pay-off to shunners, $B - C + bp(1 - e) - c[p(1 - e) + e(1 - p)]$, is greater than the pay-off to second-round defectors, $B - C + ebp$, or

$$e < \frac{p(b - c)}{2p(b - c) + c}$$

This means a population of shunners ($p = 1$) is only evolutionarily stable if the error is sufficiently small

$$e < \frac{b - c}{2b - c}$$

In other words, cooperation will unravel if second-round defectors cannot be detected most of the time. In contrast, a population of second-round defectors ($p = 0$) is stable for any positive error rate and can resist invasion even when shunners are common. Thus, the emergence of shunners when they are rare cannot be explained by the authors' model¹.

Note that the simple model presented here raises a broader concern with all models of indirect reciprocity^{3–5} and related experimental results⁶. Previous work has already shown that indirect reciprocity is stable only when donors have very reliable information about the behavioural histories of all individuals in the population^{4,7}. But even this assumes that there is no evolutionary pressure on the reliability of information. If at least some acts of giving are not observable, then individuals may have an incentive to misrepresent their behavioural histories in order to secure the benefits of indirect reciprocity without paying the costs. Considering a human context in particular, what keeps these individuals from evolving deceptive behaviours that would reduce the reliability of information and allow them to benefit from aid without providing it?

James H. Fowler

Department of Political Science, University of California at Davis, One Shields Avenue, Davis, California 95616, USA
e-mail: jhfowler@ucdavis.edu

1. Panchanathan, K. & Boyd, R. *Nature* **432**, 499–502 (2004).
2. Fehr, E. *Nature* **432**, 449–450 (2004).
3. Nowak, M. A. & Sigmund, K. *Nature* **393**, 573–577 (1998).
4. Panchanathan, K. & Boyd, R. *J. Theor. Biol.* **224**, 115–126 (2003).
5. Alexander, R. D. *The Biology of Moral Systems* (de Gruyter, New York, 1987).
6. Milinski, M., Semmann, D. & Krambeck, H. J. *Nature* **415**, 424–426 (2002).
7. Nowak, M. A. & Sigmund, K. *J. Theor. Biol.* **194**, 561–574 (1998).

doi:10.1038/nature04201

HUMAN COOPERATION

Panchanathan & Boyd reply

Reply to: J. H. Fowler *Nature* 437, doi:10.1038/nature04201 (2005)

We have shown that, if a system of indirect reciprocity is stable, exclusion from that system could deter collective-action cheats¹. Unlike direct punishment^{2–5}, indirect punishers benefit by avoiding donation, obviating the second-order free-rider problem. Fowler claims⁶, however, that we assume away the second-order free-rider problem, and (by adding a new error term) argues that indirect-reciprocity defectors undermine cooperation.

We find three flaws in Fowler's argument. First, we do not assume away the second-order free-rider problem. In models with direct punishment^{2–5}, the public-goods phase represents

a first-order collective action, whereas the punishment phase represents a second-order collective action. Such models are vulnerable to the second-order free-rider problem because selection favours strategies that avoid costly punishment of public-goods cheats (that is, second-order free riders). In our model¹, the public-goods phase is followed by an indirect-reciprocity phase. Because public-goods cheats are punished through exclusion from indirect reciprocity, selection favours punishers (shunners) over non-punishers (cooperators). Whereas cooperators represent second-order free riders, indirect punishment

is not susceptible to the second-order free-rider problem.

Second, Fowler's second-round defectors represent analogues of defecting strategies considered in previous models of indirect reciprocity^{7,8}. If they can avoid detection, Fowler is correct: indirect reciprocity may not be stable^{7,8}. Our model¹ assumes indirect reciprocity is stable as a point of departure for analysing a different problem. To link Fowler's claim to previous results, we present a model combining a public-goods game and an iterated indirect reciprocity game^{7,8}, considering only shunners and second-round defectors. In Fowler's model, error about individuals' intentions and knowledge about others' past behaviour are conflated in a single term.

Disentangling these, we first consider implementation errors⁸, in which individuals defect when they meant to cooperate and cooperate when they meant to defect. We omitted unintentional cooperation error in our original model¹ as it has no qualitative effect on the results. We also consider ignorance about the past behaviour of partners.

Using this model, we find that shunners can resist invasion by rare second-round defectors if

$$\left(\frac{wq(1-e)}{1-wqe} \right) b > c \quad (1)$$

where q represents the probability of knowing the standing (either good or bad) of a ran-

domly selected partner, w represents the probability that the indirect reciprocity game persists for another interaction, b is the benefit of receiving mutual aid, c is the cost of providing mutual aid, and e is the probability that a shunner defects when intending to cooperate. (Note the opposite error term, intentional defection resulting in cooperation, is absent in the condition described in equation (1).) This means that individuals are ignorant about the behaviour history of their partner with probability $1 - q$ and the expected number of interactions in the indirect reciprocity stage is $1/(1 - w)$. The bracketed expression on the left of condition (1) represents the degree of assortment.

In order for cooperation to evolve in any model, altruists must be able to channel cooperation to other altruists and withhold it from defectors. In indirect reciprocity, assortment is generated through reputation^{7,8}. If reputations are sufficiently accurate, cooperation can be evolutionarily stable. For example, if we set $e = 0$, the shunner strategy is stable if $wqb > c$. In inclusive-fitness models⁹, assortment is generated through kin-biased interaction, leading to Hamilton's rule ($rb > c$, where r is the coefficient of relatedness). In reciprocal-altruism models¹⁰, the persistence of dyads generates assortment, leading to the expression $wb > c$.

Third, finding that shunners cannot invade a population of second-round defectors,

Fowler raises a broad concern with all models of indirect reciprocity. This is based on a misunderstanding: in systems of reciprocity, both direct¹⁰ and indirect^{7,8}, randomly paired altruists can only increase in frequency if their initial frequency exceeds a threshold. Below this threshold, altruists too often interact with defectors and cooperation collapses. Fowler's finding that shunners cannot invade second-round defectors is therefore unremarkable. If kin-biased interactions are assumed, previous models of reciprocity have shown that altruists can increase when rare^{1,8,10}.

Karthik Panchanathan, Rob Boyd

Center for Behavior, Evolution, and Culture and Department of Anthropology, University of California, Los Angeles, California 90095, USA
e-mail: buddha@ucla.edu

1. Panchanathan, K. & Boyd, R. *Nature* **432**, 499–502 (2004).
2. Boyd, R. & Richerson, P. *Ethol. Sociobiol.* **13**, 171–195 (1992).
3. Gintis, H. *J. Theor. Biol.* **206**, 169–179 (2000).
4. Henrich, J. & Boyd, R. *J. Theor. Biol.* **208**, 79–89 (2001).
5. Boyd, R., Gintis, H., Bowles, S. & Richerson, P. *Proc. Natl Acad. Sci. USA* **100**, 3531–3535 (2003).
6. Fowler, J. H. *Nature* doi:10.1038/nature04201 (2005).
7. Nowak, M. & Sigmund, K. *J. Theor. Biol.* **194**, 561–574 (1998).
8. Panchanathan, K. & Boyd, R. *J. Theor. Biol.* **224**, 115–126 (2003).
9. Hamilton, W. D. *J. Theor. Biol.* **7**, 1–52 (1964).
10. Axelrod, R. & Hamilton, W. D. *Science* **211**, 1390–1396 (1981).

doi:10.1038/nature04202