# Prosocial preferences do not explain human cooperation in public-goods games

Maxwell N. Burton-Chellew[a,b] and Stuart A. West[b,1]

[a]Nuffield College, University of Oxford, Oxford OX1 1NF, United Kingdom; and [b]Department of Zoology, University of Oxford, Oxford OX1 3PS, United Kingdom

Edited* by Raghavendra Gadagkar, Indian Institute of Science, Bangalore, India, and approved November 9, 2012 (received for review June 27, 2012)

**It has become an accepted paradigm that humans have "prosocial preferences" that lead to higher levels of cooperation than those that would maximize their personal financial gain. However, the existence of prosocial preferences has been inferred post hoc from the results of economic games, rather than with direct experimental tests. Here, we test how behavior in a public-goods game is influenced by knowledge of the consequences of actions for other players. We found that (*i*) individuals cooperate at similar levels, even when they are not informed that their behavior benefits others; (*ii*) an increased awareness of how cooperation benefits others leads to a reduction, rather than an increase, in the level of cooperation; and (*iii*) cooperation can be either lower or higher than expected, depending on experimental design. Overall, these results contradict the suggested role of the prosocial preferences hypothesis and show how the complexity of human behavior can lead to misleading conclusions from controlled laboratory experiments.**

altruism | behavioral economics | black box | framing effect | reciprocity

Economic experiments have shown that people cooperate at levels higher than predicted if they were maximizing their financial gain (1–6). For example, in public-goods games, which are used as a model for collective-action problems, individuals can contribute money to produce benefits that are shared by all members of their group, including themselves. If an individual's share of the public benefit from his contribution is less than his contribution, then any individual wishing to maximize his own income should contribute zero. In contrast to this prediction, most individuals do contribute something (typically around 40–50%), and although contributions go down when play is repeated, groups continue to contribute 10–20% of their resources to the public good (4, 7).

These results have led to the argument that human behavior takes into account the welfare of others in a way that is perhaps unique in the animal world and cannot be explained by standard gene-based evolutionary theory (2–5, 8). Specifically, it is argued that individuals have prosocial preferences that lead them to help others, even when interacting with unrelated strangers, and when there is no scope of repeated interactions (2–5, 8). However, the existence of prosocial preferences has only been inferred post hoc from the existing data, rather than tested for directly, by varying among treatments the extent to which individuals can discern the consequences of their behavior for others (9–15).

Here, we directly test the prosocial preferences hypothesis by experimentally varying the information that individuals receive about the consequences of their behavior for others. We used a linear public-goods game, with groups of four players, in which contributions to the public good were multiplied by 1.6 before being shared out equally. This means that for each monetary unit (MU) contributed a player received 0.4 units back (the marginal per-capita return, MPCR), and so the strategy that maximized the financial gain of individuals was to contribute 0 MU (0% cooperation) from their endowments of 40 MU (3). We randomly determined group membership for each of 20 rounds to reduce the possibility for reciprocity while allowing for learning (16).

We carried out three versions of this game. The only difference between versions was the information that individuals received about the consequences of their behavior. This information was either standard, reduced, or increased (Table 1 and Fig. S1). In our standard-information treatment we told participants the payoff structure of the game and then, after each round of play, how much the others in their group had contributed and their own personal payoffs. This is the same level of information provided in many previous studies (3). To provide reduced information, we used a "black-box" treatment, in which participants were not even informed that they were playing with others in a group. Specifically, we told participants that they could input 0–40 virtual "coins" into a virtual black box, which would return a (nonnegative) output, determined by a mathematical function on their input. To provide an "enhanced-information" treatment, we carried out the same public-goods game but also provided individuals with a detailed breakdown of how much each other member of their group had contributed, received back from the public good, and earned in that round (Table 1 and Fig. S1).

These three treatments allowed us to test whether prosocial preferences are either necessary or sufficient to explain the higher-than-expected level of cooperation in public-goods games. In our black-box treatment, players could only learn the payoff-maximizing strategy (contribute 0 MU) through trial and error, and they did not even know that they were playing a game with other individuals, let alone that their contributions had beneficial consequences for others. Consequently, prosocial preferences cannot be invoked to explain the data. In contrast, in our enhanced-information treatment, players were still given the same information about the game payoffs, but they were better able to see that their cooperation was costly to themselves and beneficial to others. The prosocial preferences hypothesis predicts that this should either not alter cooperation levels or lead to relatively higher levels of cooperation. Furthermore, as an experimental control, we repeated all three of our above treatments but with a higher benefit from contributing to the public good, such that the strategy that would maximize financial gain was to cooperate 100% (contribute 40 MU).

## Results

In total, we had 16 sessions and used 236 participants, who all played both the black-box game and then, or previously, also participated in either the standard-information or the enhanced-information public-goods game, which they were informed was a separate experiment. Each treatment was played for 20 successive rounds. Overall, there was a significant difference in the mean

Author contributions: M.N.B.-C. and S.A.W. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: stuart.west@zoo.ox.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1210960110/-/DCSupplemental.

**Table 1. Comparison of the different information contents in the postdecision feedback of the three treatments**

| Treatment | Own decision | Knowledge of playing with others | Decisions of others (within-group) | Earnings of others (within-group) | Own earnings |
|---|---|---|---|---|---|
| Black box | Yes | No | No | No | Yes |
| Standard information* | Yes | Yes | Yes | No | Yes |
| Enhanced information | Yes | Yes | Yes | Yes | Yes |

*This is the same information as provided in ref. 3.

level of cooperation among these three treatments [linear mixed model (LMM): $F_{2,13.2} = 4.0$, $P = 0.044$]. The results below are on all of the data; however, we find the same qualitative results when analyzing the data from naïve participants only (*SI Text*; Figs. S2 and S3).

**Variable Information.** The level of cooperation over time was strikingly similar for the black-box treatment and the standard-information public-goods game. We found no significant difference in the mean level of cooperation between our standard-information public-goods game and the black-box treatment, with an overall mean of 9.8 MU (25%) in both [Fig. 1*A*; stepwise deletion to minimum adequate model, deviance test: $\chi^2_{(1)} < 0.1$, $P = 0.975$]. In both treatments, the mean cooperation started in the typical range of 15–20 MU (40–50%), although it was slightly lower in the black-box treatment [$\chi^2_{(1)} = 7.4$, $P = 0.007$]. By the final round, the level of cooperation was not significantly different between these two treatments and had declined to 6.3 MU [16%; $\chi^2_{(1)} < 0.1$, $P = 0.999$]. Furthermore, in the final round of play, the percentage of individuals who contributed fully was not different across the two treatments, being ~4% in both (Fisher's exact test: $P = 0.580$, Fig. S4).

In contrast, we found that when we gave individuals enhanced information about the earnings made by other players, this led to significantly lower cooperation than in the other treatments [Fig. 1*A*; $\chi^2_{(1)} = 6.1$, $P = 0.013$; round one was excluded because the difference between treatments is in the feedback given, and so treatment effects will not operate until after round one]. In both the standard- and enhanced-information treatments, individuals started cooperating at the same intermediate level of 19.5 MU [49%; $\chi^2_{(1)} < 0.1$, $P = 0.999$]. However, when we gave information detailing other people's earnings, this led to 32% less cooperation in the 19 subsequent rounds (mean of 6.7 MU versus 9.8 MU for standard information). This was driven by the rate of cooperation in the enhanced-information treatment declining approximately twice as fast in the opening rounds of the game [$\chi^2_{(1)} = 7.2$, $P = 0.007$; rate of decline in the first five rounds, −1.2 and −2.5 MU per round for standard- and enhanced-information treatments, respectively].

Our conclusions are robust to different forms of analysis. To increase our statistical power, our above results here are on all of the data pooled together. However, we reached the same qualitative conclusions when analyzing just the decisions from the first game that each player played in the session (*SI Text*) and using nonparametric Mann-Whitney *U* tests and Kolmogorov-Smirnov tests (KS) to compare the individual decisions by treatment in each time period (*SI Text*).

**Profitable Controls.** As an experimental control, we repeated all three of our above treatments using the above participants and counterbalancing treatment orders (Table S1) but with a higher benefit from contributing to the public good such that the strategy that would maximize financial gain would be to cooperate 100% (40 MU) (17). This is necessary, because the strategy that would maximize financial gain in a typical linear public-goods game is to contribute 0 MU (0%), which means that imperfect behavior and higher-than-expected cooperation will be conflated, with both leading to >0% cooperation (17–19). Specifically, we multiplied each contributed MU by 6.4 before sharing them out equally, such that for each unit contributed a player received 1.6 units back. The two conditions, costly and profitable, were thus symmetrical with regard to the percentage cost and benefit directly received by any contributing participant (60% loss or 60% gain).

We found in our profitable public-goods game that providing information about the earnings made by other players again led to a lower mean level of cooperation [Fig. 1*B*; $\chi^2_{(1)} = 4.5$, $P = 0.034$]. This was despite the fact that the initial level of cooperation in round one did not differ significantly between the standard-information and enhanced-information treatments [~25 MU, 63% of MU contributed; $\chi^2_{(1)} = 0.4$, $P = 0.511$]. The level of cooperation increased over the first five rounds in the standard-information treatment (LMM: $F_{1,7.3} = 7.6$, $P = 0.27$), but not in the enhanced-information treatment (LMM: $F_{1,7.5} < 0.1$, $P = 0.923$). Consequently, knowledge about the earnings of others not only led to a lower level of cooperation [$\chi^2_{(1)} = 13.1$, $P < 0.001$], but this effect was sufficient to prevent the contribution from increasing from a starting point of ~25 MU (63%) over the course of the game.

Considering the black-box treatment, the initial level of cooperation in round one was significantly lower [~13 MU; $\chi^2_{(1)} = 34.42$, $P < 0.001$] than in the other two treatments. After this, the level of cooperation increased in the black-box treatment (LMM: $F_{1,25.2} = 61.4$, $P < 0.001$) at a rate that did not differ significantly from that of the standard-information treatment [$\chi^2_{(1)} = 2.6$, $P = 0.105$]. As above, we reached the same qualitative conclusions when analyzing our data with nonparametric tests (*SI Text*).

Overall, combining both the costly and profitable public-goods games our results showed that an increased knowledge of the consequences of cooperation for others leads to lower overall levels of cooperation (Fig. 1; LMM: $F_{1,14} = 6.4$, $P = 0.024$), and this does not depend on whether cooperation is costly or profitable (LMM: $F_{1,14} = 0.9$, $P = 0.364$).

## Discussion

We tested how cooperation varied in a public-goods game depending on the information provided to players about the game. We found that: (*i*) the level of cooperation did not differ between a standard public-goods game and our black-box treatment, in which players did not even know their choices affected others (Fig. 1*A*); and (*ii*) providing players with enhanced information about the earnings of their groupmates led to lower levels of cooperation (Fig. 1*A*), even when cooperation was directly profitable (Fig. 1*B*). Furthermore, when we made 100% cooperation the income-maximizing and prosocial strategy, we found that cooperation failed to reach 100% in any treatment (Fig. 1*B*).

**Prosocial Preferences and a New Null Hypothesis.** Our results contradict the predictions of the prosocial preferences hypothesis in three ways. First, we found that the mean and final levels of

EVOLUTION

ECONOMIC SCIENCES

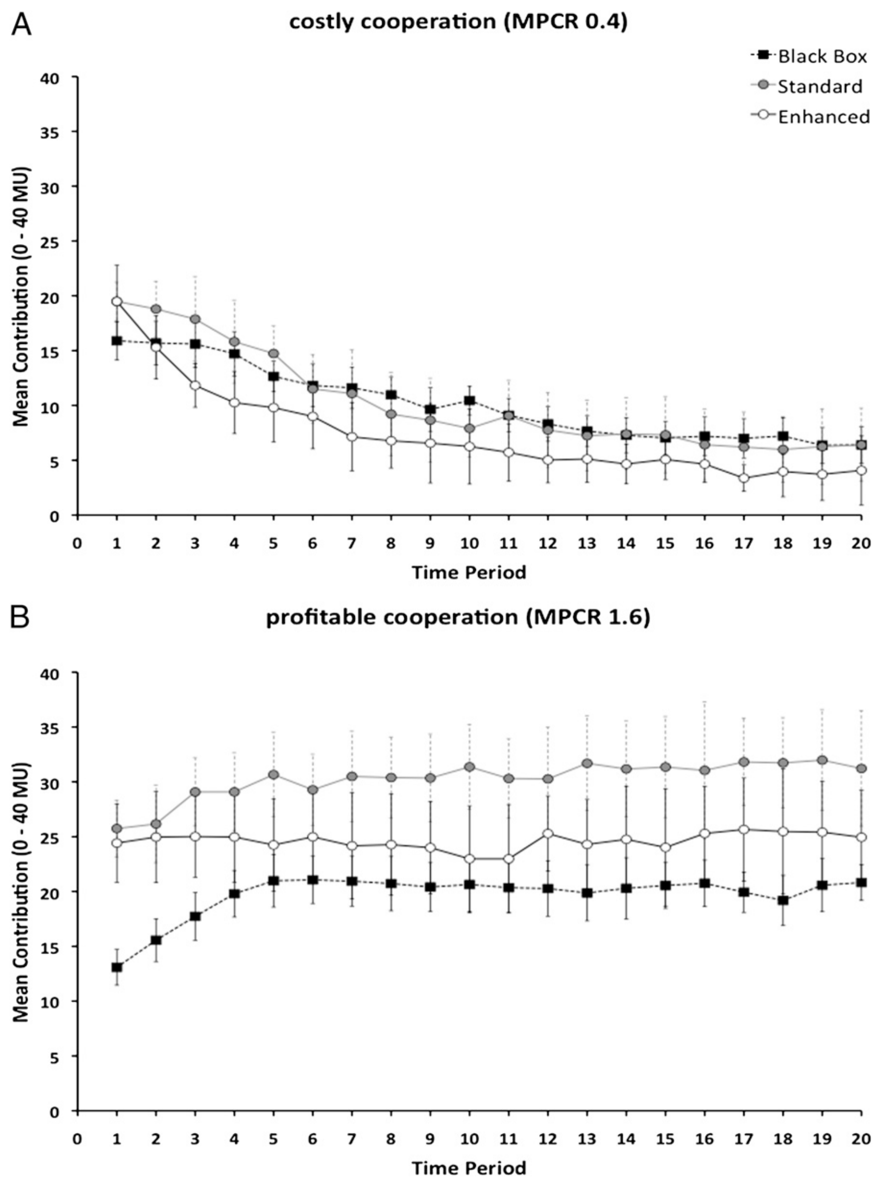


**Fig. 1.** Cooperation with different costs and different levels of information. Shown are the mean (± 95% confidence interval) cooperation levels for public-goods games per time period of the game in which cooperation was either (*A*) directly costly (return to the individual from each unit invested in the public good = 0.4, group return = 1.6) or (*B*) directly profitable (individual return = 1.6, group return = 6.4). We varied the information given to players such that they were either playing a standard-information public-goods game in which they were informed about the payoff structure of the game and the decisions of their group mates after each round (●), an enhanced-information public-goods game in which they were also informed of the earnings of their group mates (○), or a black-box game in which they were not informed that they were playing a game with others or of the decisions or earnings of others (■). Each game was played 20 times, with groups randomly recreated each time.

cooperation did not differ significantly between our standard-information public-goods game and our black-box game. The participants in our black-box game did not know they were interacting with others, or that their behavior could help others. Instead, they only had information about their own payoffs to learn from. Consequently, this shows similar levels of cooperation, and a similar decline over time, when there was no opportunity for prosocial preferences, and individuals were presumably merely trying to maximize their personal financial gain. Behavior in the standard-information public-goods game is therefore consistent with uncertain individuals who are learning the best way to maximize their personal financial gain and prosocial preferences are not necessary to explain the results.

One way of conceptualizing our black-box treatment is that it provides and tests a new experimental null hypothesis for behavior in public good games. The null model is that participants enter the game uncertain of what the best decision is to maximize their earnings, and that they are largely indifferent to the welfare of others and operate with a myopic regard for their own welfare (20). Unsurprisingly, we observed that initial contributions were intermediate, and that payoff-based learning led to an improvement over time. The prosocial preferences hypothesis makes no predictions about how people will play the black-box treatment. However, the point here is that the lack of a significant difference between behavior in our standard and black-box treatments shows that behavior in the standard game is consistent with our null hypothesis, shifting the burden of proof toward demonstrating prosocial preferences directly.

Second, we found that an increased awareness of how cooperation benefits others leads to a reduction in the level of cooperation. The prosocial preferences hypothesis would suggest that increased information would either not influence or would increase the level of cooperation. If players' preferences are being measured by the consequences for others resulting from their decisions, then this assumes that they are aware of the effects of their decisions. Therefore our increased information should have no effect on contribution levels as it is only telling players what they already know (2). Alternatively, if the players are perhaps incapable of calculating the consequences of their behavior for others, but do cooperate because of prosocial preferences, then this information would lead to increased cooperation levels. Consequently, not only do our results fail to provide experimental support for prosocial preferences, they suggest the opposite: participants cooperate less when they have clearer information about how their cooperation benefits others. Instead, the observed pattern again supports the null hypothesis, because the increased information always suggests that contributions are costly and thus may influence uncertain players to play correctly in the costly game and play incorrectly in the profitable game (from an income-maximizing perspective).

Third, prosocial preferences cannot explain our result that, in our profitable public-goods game treatments, the mean contributions in all rounds remained substantially below 40 MU (100% cooperation; Fig. 1*B*). The prosocial preferences hypothesis predicts 100% cooperation (40 MU) in this profitable game. Alternatively, it might be argued that rational individuals prioritizing within-group success could predict <100% cooperation in profitable games (Fig. 1*B*). However this would require participants to value their relative success within a temporary, transient group (we changed our group compositions randomly each round) more than their relative success within the session or the experiment as a whole (those who contribute more make more in the session). The general point here is that because previous experiments lacked appropriate controls for imperfect behavior, any imperfect behavior would have led to a higher-than-expected level of cooperation, and hence a biased conclusion.

**Alternative Explanations.** Could our results have been influenced by the potential for reciprocity (16)? Our participants played for 20 rounds and thus could expect some repeated interactions. However, this is likely to have had negligible influence for three reasons. First, group formation was random and anonymous, such that there was no potential for reputational consequences and one could not predict when a repeated interaction would occur. Second, there was the same amount of repeated interaction in all treatments, and hence this could not explain the between-treatment differences upon which we focus. Third, our data follow the normal, stylized patterns of public-goods games, either with or without repeated interactions, and it has been shown elsewhere that with random group composition there was no significant influence of whether there could be some repeated interactions (21).

Could it be argued that the black-box treatment is sufficiently similar to a social interaction to trigger any purported human tendencies to altruistically help others, and that this explains their pattern of monetary contributions? Such an explanation could explain the previous result that people "cooperate" when knowingly playing public-goods games with a computer (22). However, our black-box game was deliberately devoid of all social cues and frames, and so it is harder to envisage how this could have engaged social tendencies. Furthermore, if we accept that people cooperate maladaptively with computers because of hard-wired tendencies to be prosocial, then we would have to apply the same logic to all cooperation in economic games. This would mean that the cooperation observed in public-goods games would be explained by hard-wired responses to evolutionary factors such as the consequences of repeated interactions and reputational concerns, even when these are removed by the experimenter, and thus the prosocial preferences explanation for costly cooperation with strangers is invalidated either way (15, 16).

Could it be argued that some form of social preference, such as reciprocity or inequality aversion, can explain both the decline in cooperation over time in costly public-goods games and the lower cooperation in the enhanced-information treatment (Fig. 1*A*)? The idea here is that both of these mechanisms would place a negative value on increased earnings by others, and increased information would lead to individuals' cooperating less to either punish nonreciprocators or reduce the inequality (2, 23, 24). However, these hypotheses are clearly falsified by the fact that in our profitable control treatment the mean level of cooperation increased over time, and from an intermediate start that is not predicted by such preferences. Furthermore: (*i*) as described above, our experiment was designed to minimize the potential for reciprocity; (*ii*) as discussed in greater detail below, the inequality aversion hypothesis often implicitly assumes that individuals compute the financial consequences of their strategy from the game structure—in this case, the enhanced-information would provide no new insight, and so there should be no influence of enhanced-information (2); and (*iii*) our black-box treatment shows that the assumption that the decay in public-goods games is due to "punishing" others or reducing inequality is not required, because this decline still occurs when participants can only be responding to their payoffs (Fig. 1*A*).

**Implications.** Given that the prosocial preferences hypothesis appears neither a necessary nor sufficient explanation, how can we explain the higher-than-expected level of cooperation that is often observed in economic games? The most parsimonious explanation for our results is our null hypothesis presented above, that (*i*) individuals are trying to maximize their financial gain, but (*ii*) behavior is imperfect due to uncertainty or false beliefs, or subject to some sort of noise, which could result from a variety of factors, including errors, boredom, learning, exploration, fluctuating preferences, or evolutionary constraints. This combination of factors can explain three results that are inconsistent with prosocial preferences: (*i*) higher-than-expected levels of cooperation were seen when the income-maximizing strategy was to contribute zero (0%; Fig. 1*A*), but lower-than-expected levels of cooperation were seen when the income-maximizing strategy was to cooperate completely (100%; Fig. 1*B*) (17); (*ii*) the level of cooperation in a standard-information public-goods game did not differ significantly from that in a black-box game in which individuals had no information about the consequences of their behavior for others (Fig. 1*A*); and (*iii*) enhanced information about the earnings of others, which suggests that contributions are costly, led to lower, not higher, levels of cooperation regardless of whether cooperation was costly or profitable (Fig. 1). Although we have not directly tested the role of punishment, our results suggest the possibility of analogous problems for work on punishment, because this previous work has also assumed that the higher-than-expected levels of cooperation were because of prosocial preferences (3).

More generally, our results show how the complexity of human behavior can lead to potentially misleading conclusions from controlled laboratory settings. Even though there is evidence to suggest that many people fail to fully understand public-goods games (17, 22, 25), rational choice theory has been extended to include prosocial preferences in a post hoc attempt to explain the data (2, 8, 26): "Expanding the domain of preferences to include the utility of others provides a coherent way to extend rational choice theory" (ref. 26, abstract). Thus, the paradigm of prosocial preferences has been based on the assumption that people act rationally (perfectly) in accord with their desires and the financial consequences of their decisions (2, 8, 26). In the extreme, it has been argued that "it is not necessary for subjects to be informed about the final monetary payoffs of other subjects" and that all that is required for relative payoffs to determine behavior is knowledge of the game structure such that individuals can "compute the distributional implications" (ref. 2, footnote 6). However, our results show that humans do not behave in such a "robotlike" way and thus may not perfectly express their preferences, especially when these preferences are "measured" by the social consequences (relative payoffs) of their actions in one-shot experimental games (6, 13, 15, 27, 28). Furthermore, the level of cooperation observed is "irrationally" sensitive to parameters that do not change the rational strategy to fully defect, such as group size, group comparisons, and the presence of eyelike images (29–33). Whereas such behavior can be beneficial in the real world (15, 16, 34), it can lead to seemingly irrational behavior in experimental settings, where normally useful cues of success can be misleading, and the influence of usually important factors has been experimentally excluded. This does not mean that humans lack a sense of fairness, but rather that it is most favored when it provides a greater net benefit.

## Methods

**Participants and Sessions.** A total of 236 voluntary participants (108 females, 111 males, 17 unknown) took part in one session each. Each session lasted ∼50–60 min. We had 12 or 16 participants for each session (5 and 11 sessions, respectively) and conducted all 16 sessions at the Centre for Experimental Social Sciences (CESS) at Nuffield College, University of Oxford. CESS recruited the participants using the Online Recruitment System for Economic Experiments (35) and upon our request only recruited participants that had never before participated in a public-goods experiment (at CESS). Although we did not fully inform participants about the consequences of their decisions in the black-box treatments, we did not lie to them, and the CESS ethical committee, which forbids deception in economic experiments, approved the experiment. We programmed and conducted the experiment in z-Tree (36). The decisions and responses of the participants were anonymous, as were their earnings and payments, which were administered by the CESS administrators, who were not directly involved in the experiments. In total, we gave each participant a sum total endowment of £4.80 to allocate to either her private account or her group account/black box. The mean monetary reward was £12.40 and ranged from £6.20 to £15.50.

**Different Treatments.** An overview of the information we provided as feedback during the treatments is shown in Fig. S1. From the participants' points of view, they played two separate experiments (the black-box experiment and the public-goods experiment), each of them under two conditions for 20 times apiece. The two conditions were (*i*) cooperation is directly costly and (*ii*) cooperation is directly profitable, although participants were not necessarily aware of this distinction. The order of the treatments was counterbalanced across sessions, but the two black-box games (costly and profitable) were always the first two or the final two games of the session. Within this constraint, we used all possible treatment order permutations ($n = 2 \times 2 \times 2 = 8$) to produce an orthogonal experimental design (Table S1).

**Black-box experiment.** All participants played the same two versions of the black-box game. In short, we told the participants that they had received 40 virtual coins, worth real money, and that they had to decide to either keep all 40 coins or to input a fraction (0–40 coins) of them into a "black box." They then read the following information (a full copy of the instructions is given in *SI Appendix*): "This 'black box' performs a mathematical function that converts the number of coins inputted into a number of coins to be outputted. The function contains a random component, so if two people were to put the same amount of coins into the 'black box' they would not necessarily get the same output. The number outputted may be more or less than the number you put in, but it will never be a negative number, so the lowest outcome possible is to get 0 (zero) back. If you choose to input 0 (zero) coins, you may still get some back from the box."

We also told them that they would play for 20 turns, receiving a new set of 40 coins each time, and that the most they could ever input was 40 coins. We also explained that after their 20 turns they would start again with a new, potentially different black box for another 20 turns. In addition, we explained that their final income from each turn would be their initial 40 coins, minus their input, plus all of the coins, if any, that they get back. The feedback we gave them after each turn comprised four items: "initial coins" (i.e., their endowment), "input" (i.e., their contribution), "output returned" (i.e., their returns from the group project), and their final balance from that round (Fig. S1).

**Statistical Analysis.** Because the responses of participants within a session are nonindependent we used the mean level of contributions for each session per time period to avoid pseudoreplication. We used LMM and fitted "session ID" as a random factor to account for the repeated measures over time. When comparing the mean levels of contributions for the whole treatment, we excluded the first time period of each treatment. This was because the treatment effect was not applied until after the first contributions had been made, and thus any differences in the first time period could only be due to chance. We compared the relative merit of different models by using maximum likelihood models and comparisons of the residual deviance. The comparison is made with a $\chi^2$ difference test of the change in deviance between models, with the degrees of freedom set by the difference in the number of parameters between two nested models. Models are progressed from the null model through to the maximal model, and vice versa (although the order is unimportant with orthogonal experimental designs such as ours), in search of the minimal adequate model (MAM). The MAM is the model with the smallest residual deviance for a given level of parameters and is identified on the basis of deletion/addition tests. These are $\chi^2$ tests that assess the significance of the change in deviance that results when a given term is removed/inserted. The MAM is the model that can no longer be improved by the insertion of additional terms or factor levels (37).

**ACKNOWLEDGMENTS.** We thank Luis Miller at the Centre for Experimental Social Sciences for recruiting participants, the reviewers for their helpful comments, Nuffield College for facilities, and the European Research Council for funding.

1. Camerer CF, Fehr E (2006) When does "economic man" dominate social behavior? *Science* 311(5757):47–52.
2. Fehr E, Schmidt KM (1999) A theory of fairness, competition, and cooperation. *Q J Econ* 114(3):817–868.
3. Fehr E, Gächter S (2002) Altruistic punishment in humans. *Nature* 415(6868):137–140.
4. Fehr E, Fischbacher U (2003) The nature of human altruism. *Nature* 425(6960): 785–791.
5. Gintis H, Bowles S, Boyd R, Fehr E (2003) Explaining altruistic behavior in humans. *Evol Hum Behav* 24(3):153–172.
6. Henrich J, et al. (2005) "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behav Brain Sci* 28(6):795–815, discussion 815–755.
7. Ledyard J (1995) Public goods: A survey of experimental research. *Handbook of Experimental Economics*, eds Kagel J, Roth A (Princeton Univ Press, Princeton), pp 253–279.
8. Fehr E, Fischbacher U (2005) Human altruism: Proximate patterns and evolutionary origins. *Anal Kritik* 27:6–47.
9. Binmore K, Shaked A (2010) Experimental economics: Where next? *J Econ Behav Organ* 73(1):87–100.
10. Blanco M, Engelmann D, Normann HT (2011) A within-subject analysis of other-regarding preferences. *Games Econ Behav* 72(2):321–338.
11. Burnham TC, Johnson DDP (2005) The biological and evolutionary logic of human cooperation. *Anal Kritik* 27:113–135.
12. Harrison GW (2008) Neuroeconomics: A critical reconsideration. *Econ Philos* 24(3): 303–344.
13. Oechssler J (2003) Intentions matter: Lessons from bargaining experiments — Comment on, reasons for conflict: Lessons from bargaining experiments, by Falk, A., Fehr, E., and Fischbacher, U. *J Inst Theor Econ* 159(1):195–198.
14. Ross D (2012) What can economics contribute to the study of human evolution? *Biol Philos* 27:287–297.
15. Trivers RL (2004) Mutual benefits at all levels of life. *Science* 304(5673):964–965.
16. Trivers RL (1971) Evolution of reciprocal altruism. *Q Rev Biol* 46(1):35.
17. Kümmerli R, Burton-Chellew MN, Ross-Gillespie A, West SA (2010) Resistance to extreme strategies, rather than prosocial preferences, can explain human cooperation in public goods games. *Proc Natl Acad Sci USA* 107(22):10125–10130.
18. Hagen EH, Hammerstein P (2006) Game theory and human evolution: A critique of some recent interpretations of experimental games. *Theor Popul Biol* 69(3):339–348.
19. Engelmann D, Strobel M (2010) Inequality aversion and reciprocity in moonlighting games. *Games* 1:459–477.
20. Young HP (2009) Learning by trial and error. *Games Econ Behav* 65(2):626–643.
21. Fehr E, Gachter S (2000) Cooperation and punishment in public goods experiments. *Am Econ Rev* 90(4):980–994.
22. Houser D, Kurzban R (2002) Revisiting kindness and confusion in public goods experiments. *Am Econ Rev* 92(4):1062–1069.
23. Levine DK (1998) Modeling altruism and spitefulness in experiments. *Rev Econ Dyn* 1(3):593–622.
24. Fehr E, Gachter S (2000) Fairness and retaliation: The economics of reciprocity. *J Econ Perspect* 14(3):159–181.
25. Andreoni J (1995) Cooperation in public-goods experiments: Kindness or confusion? *Am Econ Rev* 85(4):891–904.

26. Sobel J (2005) Interdependent preferences and reciprocity. *J Econ Lit* 43(2):392–436.
27. Haselton MG, Nettle D (2006) The paranoid optimist: An integrative evolutionary model of cognitive biases. *Pers Soc Psychol Rev* 10(1):47–66.
28. McCullough ME, Kimeldorf MB, Cohen AD (2008) An adaptation for altruism? The social causes, social effects, and social evolution of gratitude. *Curr Dir Psychol Sci* 17(4):281–285.
29. Isaac RM, Walker JM (1988) Group-size effects in public-goods provision: The voluntary contributions mechanism. *Q J Econ* 103(1):179–199.
30. Burton-Chellew MN, West SA (2012) Pseudocompetition among groups increases human cooperation in a public-goods game. *Anim Behav* 84:947–952.
31. Haley KJ, Fessler DMT (2005) Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evol Hum Behav* 26(3):245–256.
32. Bateson M, Nettle D, Roberts G (2006) Cues of being watched enhance cooperation in a real-world setting. *Biol Lett* 2(3):412–414.
33. Burnham TC, Hare B (2007) Engineering human cooperation: Does involuntary neural activation increase public goods contributions? *Hum Nature-Int Bios* 18(2):88–108.
34. Trivers RL (2006) Reciprocal altruism: 30 years later. *Cooperation in Primates and Humans: Mechanisms and Evolution*, eds Kappeler PM, van Schaik CP (Springer, Berlin).
35. Greiner B (2004) The online recruitment system ORSEE 2.0: A guide for the organization of experiments in economics. *Working Paper Series in Economics* 10 (Department of Economics, University of Cologne, Cologne, Germany).
36. Fischbacher U (2007) z-Tree: Zurich toolbox for ready-made economic experiments. *Exp Econ* 10(2):171–178.
37. Crawley MJ (2007) *The R Book* (Wiley, Chichester, UK), p viii.

# Supporting Information

Burton-Chellew and West 10.1073/pnas.1210960110

## SI Results

Fig. S4 *A* and *B* shows the percentage of "free riders" and "cooperators" over time for each treatment. Although Fig. S4*A* shows that there were more extreme free riders (contributing 0 MU) in the public-goods games than in the black-box game, this difference is no longer present when free riders are defined as contributing 0–10 MU (0–25%). This pattern is largely because the participants playing the black-box game were less inclined to input 0 MU even when they had learned that low contributions were most profitable.

**Experienced vs. Naïve Participants.** Fig. S2 reveals that there was no difference in how participants played the costly black-box game depending on whether their session had just played the public-goods games beforehand or not [linear mixed model (LMM): $F_{1,14} = 0.1$, $P = 0.772$]. None of the participants had ever participated in a public-goods game at the CESS laboratory before our experiment.

When only analyzing the data generated from participants who were naïve with respect to the other game (i.e., those who if they were playing the black-box game had never played the public-goods game, and vice versa), we found results that were qualitatively similar to those in the main text (Fig. S3). Although with these data our statistical power was reduced to just four sessions each of the public-goods games (standard information and enhanced information) and eight sessions of the black-box game, we still found a significant reduction in cooperation when participants received the enhanced information compared with the standard information (LMM: $F_{1,6} = 8.3$, $P = 0.028$) and compared with the black-box game (LMM: $F_{1,10} = 5.4$, $P = 0.043$). In contrast, there was no significant difference between the black-box game and the standard-information treatment (LMM: $F_{1,10} = 2.1$, $P = 0.177$).

**Nonparametric Tests.** Still only using the data from participants who were naïve about the other game, we also used nonparametric Mann-Whitney *U* tests and Kolmogorov-Smirnov tests (KS) to compare the individual decisions by treatment in each time period and thus gain more statistical power. We found the same qualitative results as presented in the main text. These analyses further confirmed the damaging effect of the enhanced information; participants in the enhanced-information treatment were found to be less cooperative than those given the standard information in all 19 (18 KS) of the 19 rounds after round one. They also confirmed the similarity between the black-box and the standard-information treatments, which differed in only one (five KS) of the 19 repeated rounds. We can also confirm that this lack of difference was not due to a lack of statistical power to compare the two games, because participants in the black-box treatment were significantly more cooperative in 15 (12 KS) of the 19 rounds compared with participants in the enhanced-information treatment.

We also used nonparametric tests for the profitable controls and found the same qualitative results. These analyses again further confirmed the damaging effect of the enhanced information; players were found to be less cooperative than those given standard information in 14 (14 KS) of the 19 rounds after round one. However, participants in the black-box treatment were less cooperative in 19 (19 KS) and 12 (13 KS) of the 19 rounds compared with those who participated in the standard-information and the enhanced-information treatments, respectively.

## SI Discussion

The level of information in our standard and enhanced-information treatments is, although different, also equivocal. This is because the information we present in the enhanced-information game on the earnings of groupmates can be computed from the distribution of decisions, which is presented in both treatments. Consequently, the model of a rational but prosocial actor that has been suggested would predict no difference in behavior between these two treatments (1–3). In addition, any behavior that is postulated to be a response to the decisions of a participant's groupmates should be equally predicted in both treatments and therefore cannot explain the difference between our treatments. For example, any reduction in cooperation in response to the cooperation of others should be the same for the two treatments, be it a postulated "social" response to "free riders" or a postulated "antisocial" response to "high cooperators." The equivalence of our treatments also means that the fact that repeated interactions were not entirely ruled out in our treatments should not affect one treatment more or less than another, and any qualitative comparisons still stand.

The focus of our study was a qualitative comparison of different treatments. The results of economic games can be interpreted with either a qualitative or a quantitative approach (4). The qualitative approach examines differences in behavior between treatments, which typically have different payoff structures for the game that is being played, but may, as in our case, just be "framed" differently. In contrast, the quantitative approach compares the quantitative level of behavior within a specific game or treatment with that predicted by theoretical models (5–7). The deductions from this quantitative approach typically rely on the implicit assumption that individuals are rational and fully aware of all of the consequences (for themselves and for others) of their decisions (2, 8). The quantitative approach can make no predictions for a difference between our standard and enhanced-information treatments.

This does not mean that psychological concerns and emotions are not driving people's responses to our games. Rather, it means that to assume a set of a participant's preferences based on the payoff consequences of his or her decision for both him- or herself and others is unlikely to be correct. For example, as suggested by the comparison of our black-box and standard-information treatments, participants may only be trying to affect their own payoffs, and their effect on others is a mere byproduct. Alternatively, participants may be responding emotionally to the payoffs of others, as in our enhanced-information treatment, but not the decisions of others, even those these equate to the same information. This is suggested by a comparison of our standard- and enhanced-information treatments, in which case participants can no longer be assumed to have been responding to the earnings of others in the standard-information treatment, which is the same level of information as presented in ref. 9.

In summary, if participants are responding to the decisions of others, there should be no difference between our standard- and enhanced-information treatments, whether costly or profitable. Conversely, if participants are reacting to the presented information on payoffs in our enhanced-information treatment, then previous explanations that assumed participants calculated such earnings from the distribution of decisions and acted accordingly to reduce inequity are false.

Finally, it has been observed that people contribute more in public-goods games when the benefit to others is increased. This is because contributions are, on average over time, higher when the marginal per-capita return (MPCR) is higher (10). It might be argued that this suggests higher cooperation when the benefits to others are higher, implying a concern for others. However, this pattern can also be explained by the fact that a higher MPCR results in cooperation's being cheaper, the losses incurred from incorrect strategies (from an income-maximizing point of view) being lesser, and the scope for learning over time being reduced. Learning is impeded because as the MPCR tends toward 1.0, the effect one's contributions have on one's own income diminishes, and so payoffs are increasingly random with respect to one's own decisions.

## SI Methods

Table S1 details the order of treatments by session. Fig. S1 shows a graphical display of the different information we gave to our participants per treatment. *SI Appendix* is a complete copy of the black-box instructions.

## SI Appendix

**Black-Box Instructions.** Participants received the following on-screen instructions (in z-Tree) at the start of the Black Box game and had to click an on-screen button saying "I confirm I understand the instructions" before the game would begin.

***Instructions.*** Welcome to the experiment. You have been given 40 virtual coins. Each '*coin'* is worth real money. You are going to make a decision regarding the investment of these 'coins'. This decision may increase or decrease the number of 'coins' you have. The more 'coins' you have at the end of the experiment, the more money you will receive at the end.

During the experiment we shall not speak of £ Pounds or Pence but rather of "Coins". During the experiment your entire earnings will be calculated in Coins. At the end of the experiment the total amount of Coins you have earned will be converted to Pence at the following rate: 100 Coins = 15 Pence. In total, each person today will be given 3,200 coins (£4.80) with which to make decisions over two economic experiments and their final totals, which may go up or down, will depend on these decisions.

***Decision.*** You can choose to keep your coins (in which case they will be 'banked' into your private account, which you will receive at the end of the experiment), or you can choose to put some or all of them into a *'black box'.*

This '*black box*' performs a mathematical function that converts the number of coins inputted into a number of coins to be outputted. The function contains a random component, so if two people were to put the same amount of coins into the '*black box*', they would not necessarily get the same output. The number outputted may be more or less than the number you put in, but it will never be a negative number, so the lowest outcome possible is to get 0 (zero) back. If you choose to input 0 (zero) coins, you may still get some back from the box.

Any coins outputted will also be 'banked' and go into your private account. So, your final income will be the initial 40 coins, minus any you put into the '*black box*', plus all of the coins you get back from the '*black box*'.

You will play this game 20 times. Each time you will be given a **new set** of 40 coins to use. Each game is separate but the '*black box*' remains the same. This means you cannot play with money gained from previous turns, and the maximum you can ever put into the '*black box*' will be 40 coins. And you will never run out of money to play with as you get a new set of coins for each go. The mathematical function will not change over time, so it is the same for all 20 turns. However, as the function contains a random component, the output is not guaranteed to stay the same if you put the same amount in each time.

After you have finished your 20 turns, you will play one further series of 20 turns but with a new, and potentially different '*black box*'. The two boxes may or may not have the same mathematical function as each other, but the functions will always contain a random component, and the functions will always remain the same for the 20 turns. You will be told when the 20 turns are finished and it is time to play with a new black box.

If you are unsure of the rules please hold up your hand and a demonstrator will help you.

I confirm I understand the instructions.


1. Fehr E, Schmidt KM (1999) A theory of fairness, competition, and cooperation. *Q J Econ* 114(3):817–868.
2. Fehr E, Fischbacher U (2005) Human altruism: Proximate patterns and evolutionary origins. *Anal Kritik* 27:6–47.
3. Sobel J (2005) Interdependent preferences and reciprocity. *J Econ Lit* 43(2):392–436.
4. Davies NB, Krebs JR, West SA (2012) *An Introduction to Behavioural Ecology* (Wiley-Blackwell, Oxford), 4th Ed, p xii.
5. Levine DK (1998) Modeling altruism and spitefulness in experiments. *Rev Econ Dyn* 1(3):593–622.
6. Fehr E, Fischbacher U (2003) The nature of human altruism. *Nature* 425(6960):785–791.
7. Camerer CF, Fehr E (2006) When does "economic man" dominate social behavior? *Science* 311(5757):47–52.
8. Dufwenberg M, Gachter S, Hennig-Schmidt H (2011) The framing of games and the psychology of play. *Games Econ Behav* 73(2):459–478.
9. Fehr E, Gächter S (2002) Altruistic punishment in humans. *Nature* 415(6868):137–140.
10. Zelmer J (2003) Linear public goods experiments: A meta-analysis. *Exp Econ* 6: 299–310.

**(a)**

| GAME SUMMARY | Number of Coins |
|---|---|
| Initial Coins | 40 |
| Minus (-) your input | 15 |
| Plus (+) the output returned | 24 |
| Your final number of coins | 49 |

This screen lists your decisions and the results, along with your income (for this turn).

**(b)**

| Player Name | Contribution (0-40) |
|---|---|
| You | 15 |
| player A | 10 |
| player B | 30 |
| player C | 5 |
| Total contributions | 60 |
| Total after growth | 96 |

This screen lists the decisions of you and the other players (in random order) in your group (for this round). Remember the terms Player A, B, & C **are meaningless** as you play with randomly selected people in each round. The group's total contributions and the new total after the 'growth' stage are also shown.

**(c)**

| Player Name | Contribution (0-40) | Player Income = | Credits retained | + Credits returned | = Total credits |
|---|---|---|---|---|---|
| **You** | 15 | For you = | 25 | 24 | 49 |
| player A | 10 | player A | 30 | 24 | 54 |
| player B | 30 | player B | 10 | 24 | 34 |
| player C | 5 | player C | 35 | 24 | 59 |

Your income from your group's contributions and subsequent 'growth' is shown.

This screen lists the decisions and earnings of you and the other players (in random order) in your group (for this round).

Remember the terms Player A, B, & C **are meaningless** as you play with randomly selected people in each round.

**Fig. S1.** (*A*) The postdecision feedback information that participants received in the black-box treatment. (*B*) The first feedback screen in both the standard- and the enhanced-information treatments. (*C*) The second feedback screen, which differed between the two treatments (not shown in black-box treatment). Dashed lines border the information that was only shown in the enhanced-information treatment.

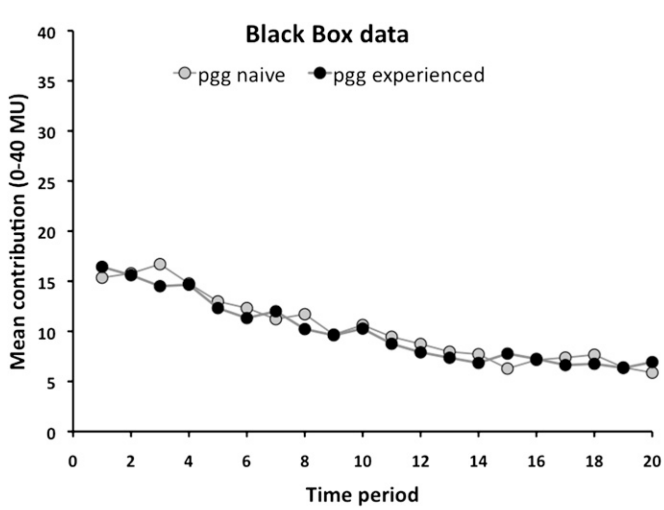


**Fig. S2.** The mean contributions into the costly black box, both when the session of participants were either naïve to public-goods games (ppg naïve, gray filled circles) or experienced in public-goods games (ppg experienced, black filled circles).

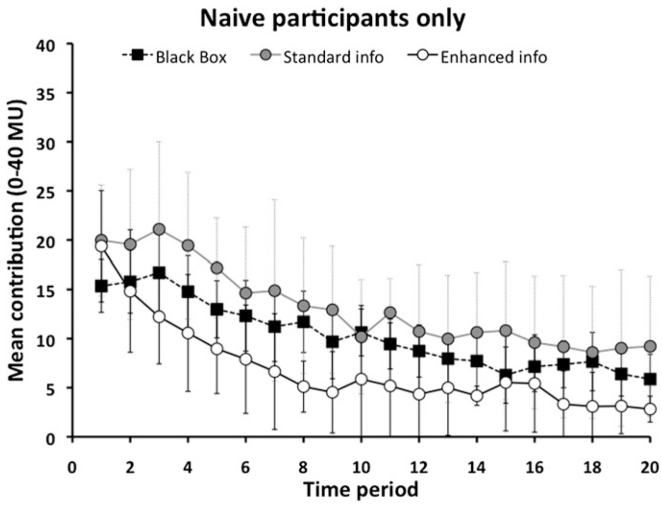


**Fig. S3.** The mean contributions in the costly public-goods game for all three treatments when the participants had not played any versions of the other game (black-box or public-goods game) beforehand ($n = 4$ for the standard-information and enhanced-information treatments and $n = 8$ for the black-box treatment).

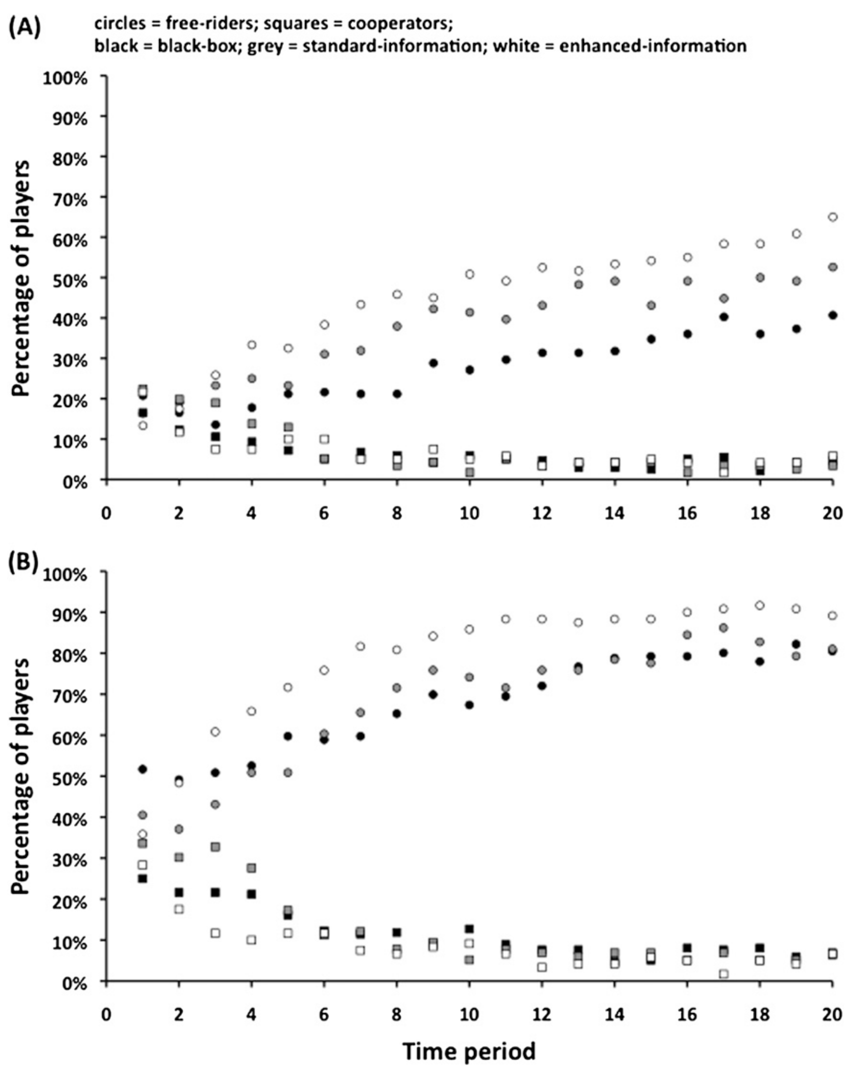


**Fig. S4.** The percentage of "cooperators" and "free riders" over time for the three different treatments in the costly public-goods game (where the return to the individual from each monetary unit invested in the public-good = 0.4 and the return to the group = 1.6). In *A*, "cooperators" are defined narrowly as those that contributed 40 MU (100%) and "free riders" are defined narrowly as those that contributed 0 MU (0%). In *B*, "cooperators" are defined more broadly as those that contributed 30–40 MU (75–100%) and "free riders" as those that contributed 0–10 MU (0–25%).

**Table S1. Experimental design detailing order of treatments for each of the 16 sessions**

| Experiment 1 (2 x 20 rounds) | | Experiment 2 (2 x 20 rounds) | |
|---|---|---|---|
| Treatment 1 | Treatment 2 | Treatment 3 | Treatment 4 |
| Black box (–) | Black box (+) | Standard (–) | Standard (+) |
| Black box (+) | Black box (–) | Standard (+) | Standard (–) |
| Black box (–) | Black box (+) | Enhanced (–) | Enhanced (+) |
| Black box (+) | Black box (–) | Enhanced (+) | Enhanced (–) |
| Black box (–) | Black box (+) | Standard (+) | Standard (–) |
| Black box (+) | Black box (–) | Standard (–) | Standard (+) |
| Black box (–) | Black box (+) | Enhanced (+) | Enhanced (–) |
| Black box (+) | Black box (–) | Enhanced (–) | Enhanced (+) |
| Standard (–) | Standard (+) | Black box (–) | Black box (+) |
| Standard (+) | Standard (–) | Black box (–) | Black box (+) |
| Standard (–) | Standard (+) | Black box (+) | Black box (–) |
| Standard (+) | Standard (–) | Black box (+) | Black box (–) |
| Enhanced (–) | Enhanced (+) | Black box (–) | Black box (+) |
| Enhanced (+) | Enhanced (–) | Black box (–) | Black box (+) |
| Enhanced (–) | Enhanced (+) | Black box (+) | Black box (–) |
| Enhanced (+) | Enhanced (–) | Black box (+) | Black box (–) |

Each row corresponds to a unique session. The (–) refers to the costly public-goods games and the (+) to the profitable versions.