Reputation-based cooperation: empirical evidence for behavioral strategies

Swakman, V.; Molleman, L.; Ule, A.; Egas, M.

DOI
10.1016/j.evolhumbehav.2015.12.001

Publication date
2016

Document Version
Final published version

Published in
Evolution and Human Behavior

License
Article 25fa Dutch Copyright Act

Link to publication

Citation for published version (APA):
Swakman, V., Molleman, L., Ule, A., & Egas, M. (2016). Reputation-based cooperation: empirical evidence for behavioral strategies. *Evolution and Human Behavior*, *37*(3), 230-235. https://doi.org/10.1016/j.evolhumbehav.2015.12.001

Evolution and Human Behavior 37 (2016) 230–235

Contents lists available at ScienceDirect

Evolution and Human Behavior

journal homepage: www.ehbonline.org

Original Article

Reputation-based cooperation: empirical evidence for behavioral strategies[☆],[☆☆]

Violet Swakman [a,1], Lucas Molleman [b,*,1], Aljaž Ule [c,d], Martijn Egas [a]

[a] Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, P.O. Box 94240, 1090 GE Amsterdam, The Netherlands
[b] The Centre for Decision Research and Experimental Economics (CeDEx), University of Nottingham, Sir Clive Granger Building, University Park, Nottingham NG7 2RD, United Kingdom
[c] Center for Research on Experimental Economics and political Decision-making (CREED), University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands
[d] Famnit, University of Primorska, Glagoljaška 8, SI-6000 Koper, Slovenia

ARTICLE INFO

Article history:
Initial receipt 21 August 2015
Final revision received 11 December 2015

Keywords:
Indirect reciprocity
Experiment
Human cooperation
Individual differences

ABSTRACT

Human cooperation in large groups can emerge when help is channeled towards individuals with a good reputation of helping others. Evolutionary models suggest that, for reputation-based cooperation to be stable, the recipient's reputation should be based not only on his past behavior (1st-order information) but also on the past behavior of the recipient's recipient (2nd-order information). Second-order information reflects the context of others' actions, and allows people to distinguish whether or not giving (or denying) help was justified. Little is known yet about how people actually condition their cooperation on 2nd-order information. With a behavioral experiment, we show that people actively seek 2nd-order information and take this into account in their own helping decisions. In an anonymous iterated helping game, donors learned if their recipients helped others in the past and could obtain 2nd-order information about these actions. Donors often requested this 2nd-order information and were especially interested to know why help was denied (i.e., defection). Justified defection was rewarded: help was generally directed towards those who defected against the selfish, and away from those who defected against helpful individuals. A detailed analysis of individual strategies reveals that many subjects based their decisions solely on 1st-order information about their recipients' past behavior. However, a substantial fraction of subjects consistently considered also the 2nd-order information about their recipients' behavior. Our results provide strong empirical support for the mechanisms that theoretically underpin reputation-based cooperation, and highlight pronounced individual variation in human cooperative strategies.

1. Introduction

Cooperation is commonly observed in the form of individuals helping others. This phenomenon is particularly puzzling when providing help is costly. For cooperation to thrive, specific mechanisms are needed to suppress 'defectors', who avoid the costs of cooperation because they do not help others. Such mechanisms either enforce cooperation through punishment (Ostrom, Walker, & Gardner, 1992; Yamagishi, 1986), or channel help towards cooperators through a form of assortment (*e.g.* kin selection (Hamilton, 1964), group selection (Maynard Smith, 1964) or partner choice (Noe & Hammerstein, 1994; Ule, 2008)). When individuals interact repeatedly, costly helping can pay off if the recipient of help returns the favor later on (Trivers, 1971). Such 'direct reciprocity' is widespread among humans (e.g. Gintis, 2000; Trivers, 1971) and is believed to be present in some animal species as well (Clutton-Brock, 2009; Dugatkin, 1997).

Human cooperation can also be supported by indirect reciprocity (Alexander, 1987; Darwin, 1871), where individuals do not base their helping on personal experience, but rather on their recipient's reputation. People can build up a good reputation by being helpful, or a bad reputation when they tend to be selfish, and reputation may spread through social networks by gossip. Providing costly help can pay off when the resulting good reputation attracts more help from others in return (Nowak & Sigmund, 2005; Sigmund, 2012). Indeed, experimental findings confirm that people help recipients more often when these recipients have been helpful towards others in the past (Seinen & Schram, 2006).

Theoretical analyses indicate that the dynamics of reputation-based cooperation critically depend on how information about past interactions gives rise to reputations (Brandt & Sigmund, 2004; Nowak & Sigmund, 1998, 2005; Ohtsuki & Iwasa, 2004). The reputation of a potential recipient of help can be based not only on information about his past helpfulness (1st-order information), but also on information about the helpfulness of the recipient's recipient (2nd-order information). Second-order information reflects the context in

[☆] **Funding:** L.M. acknowledges support by an Ubbo Emmius grant from the University of Groningen and the European Research Council (ERC-AdG295707).

[☆☆] **Author note:** The experimental data used for the analyses are available in the Online Supplementary material.

[*] Corresponding author. Tel.: +44 115 84 68088.
E-mail address: lucasmolleman@gmail.com (L. Molleman).

[1] Contributed equally to this manuscript.

http://dx.doi.org/10.1016/j.evolhumbehav.2015.12.001
1090-5138/© 2016 Elsevier Inc. All rights reserved.

which a recipient made their decision, and potentially indicates the reasons why the recipient gave or denied help. First-order information may be easier to observe than 2nd-order information, but it is not sufficient for stable cooperation; when reputations are solely based on past behavior, cooperation is likely to be destabilized by strategies that do not channel help towards helpful others, but help only to maintain their own good reputation (Leimar & Hammerstein, 2001; Nowak & Sigmund, 1998).

By contrast, stable cooperation can be supported by strategies that assess both 1st-order and 2nd-order information (Alexander, 1987; Ohtsuki & Iwasa, 2004, 2006; Sigmund, 2012). Such strategies require that the decision to help is not only based on information about a recipient's previous helpfulness, but also on information about the helpfulness of the persons to whom this recipient gave (or denied) help. One example is the 'standing' strategy that assesses a person's decision to deny help: his reputation is only damaged when he had denied help to a cooperator, but not when he had denied help to a defector (Sugden, 1986). Another example is the 'judging' strategy which in addition evaluates whether a person's decision to give help was justified: his reputation is damaged if he gave help to a defector (Nowak & Sigmund, 2005; Pacheco, Santos, & Chalub, 2006). These strategies encourage individuals to channel help towards cooperators and away from defectors, thereby stabilizing cooperation (Brandt & Sigmund, 2004; Ohtsuki & Iwasa, 2004, 2006; Panchanathan & Boyd, 2003).

Whereas 1st-order information about past helping behavior has been firmly shown to promote reputation-based cooperation in human groups (Molleman, van den Broek, & Egas, 2013; Seinen & Schram, 2006; Wedekind & Braithwaite, 2002; Wedekind & Milinski, 2000), experimental evidence for the effects of 2nd-order information is limited and conflicting. On the one hand 2nd-order information has been shown not to affect indirect reciprocity (Milinski, Semmann, Bakker, & Krambeck, 2001), and is not used very often (Ule, Schram, Riedl, & Cason, 2009) but on the other hand its availability may promote cooperation (Bolton, Katok, & Ockenfels, 2005). Crucially, there is no systematic analysis about the role of 2nd-order information in people's actual decision making process, and whether they compute standing or apply a judging strategy. When people can base their decisions on rich sets of information (as in Milinski et al., 2001; Ule et al., 2009), integration of such information may be computationally taxing (dos Santos, Braithwaite, & Wedekind, 2014; Panchanathan & Boyd, 2003). Indeed, cognitive constraints may partly explain why empirical insights into the role of 2nd-order information are limited as yet. Moreover, theory has shown that cooperation can be maintained under a large number of strategies that assess reputations in different ways, but it remains unclear whether evolution will lead to one commonly held strategy, or that a variety of privately held strategies can coexist (Brandt & Sigmund, 2005; Mashima & Takahashi, 2005; Pacheco et al., 2006; Sigmund, 2012; Takahashi & Mashima, 2006). This issue is not trivial: recent empirical work has revealed substantial individual variation in terms of strategic decision making (Engelmann & Fischbacher, 2009; Kurzban & Houser, 2005; Molleman, van den Berg, & Weissing, 2014; Ule et al., 2009), and it has been shown that this variation can strongly affect the outcome of social interactions (Fischbacher & Gachter, 2010; Gavrilets, 2015; Hartig, Irlenbusch, & Koelle, 2015; McNamara & Leimar, 2010; Molleman et al., 2014; van den Berg, Molleman, Junikka, Puurtinen, & Weissing, 2015, van den Berg, Molleman, & Weissing, 2015; Wolf, van Doorn, & Weissing, 2008). None of this has yet been explored with respect to strategies that use 2nd-order information.

Here we report on a decision-making experiment with human subjects, designed to directly examine strategies that in theory underpin reputation-based cooperation. In our experiment, subjects interacted for 100 rounds in groups of 10. Each round, the subjects were randomly and anonymously paired. In each pair, one subject was randomly assigned the role of 'donor' and the other the role of 'recipient'. Donors decided to either GIVE or DENY costly help to their recipient. Before making this decision, the donors could observe (1st-order) information about the recipient's most recent three GIVE/DENY decisions towards others. To investigate the use of 2nd-order information, we gave the donors the possibility to access information behind one GIVE/DENY decision of their recipient: a donor could observe the 1st-order information that their recipient had when they made that decision (see Supplementary Information, section S4 for screenshots, available on the journal's website at www.ehbonline.org). Together, the information available to the donors was designed to be sufficiently rich to apply strategies with key elements predicted by theory (conditioning cooperation on 1st-order and/or 2nd-order information), yet sufficiently simple to avoid cognitive overload in making decisions in our experiment.

To test how 1st- and 2nd-order information shapes reputation-based cooperation, we considered three experimental conditions. In the FREE condition (six replicate groups) 2nd-order information was available without a cost. In reality such information is often not readily available and obtaining it may require effort. In the COSTLY condition (six replicate groups) we therefore incorporated a cost for requesting 2nd-order information. In the CONTROL condition (four replicate groups), no 2nd-order information was available. We first analyze how on the aggregate level cooperation is affected by 1st-order information (previous decisions of recipients), by 2nd-order information (previous decisions of recipients' recipients), and by previous decisions of donors themselves (reflecting their current reputations). Subsequently, we zoom in on patterns of individual decision-making and examine the empirical support for strategies that in theory facilitate stable cooperation.

2. Material and methods

2.1. Experimental setup

We ran 16 replicate sessions at the CREED laboratory at the University of Amsterdam and the Sociology laboratory of the University of Groningen. Participation was by informed consent. In total, $n = 160$ subjects (80 male, 80 female, average age: 22.6 years) attended the sessions, participating in groups of 10 (CONTROL: four replicates; COSTLY: six replicates; FREE: six replicates). The experiment was conducted using z-Tree (Fischbacher, 2007); code available upon request.

Subjects interacted in 100 rounds of an 'indirect helping game'. In each round pairs were randomly formed, and within each pair the roles of donor and recipient were randomly assigned. Interactions were anonymous; subjects never knew with whom they were paired. Donors decided between two options, clicking on either 'blue' or 'purple'. With blue (GIVE help) a donor increased the earnings of his recipient by 250 points and decreased his own earnings by 200 points. With purple (DENY help) neither earnings increased or decreased (see Supplementary Information, section S4 for screenshots, available on the journal's website at www.ehbonline.org). Recipients did not have to do anything, and a waiting screen was displayed requesting to wait for the decision of the coupled donor. At the end of each round, the donor's decision and its payoff consequences were shown to the donor and the recipient. At the end of the experiment the total earnings were exchanged for money, where 300 points were worth 1 euro.

Before making their decisions, the screen of the donors always displayed the three most recent decisions of the recipient (1st-order information). These decisions were displayed in random order to avoid potential confounding effects of recency reflected in the order in which the recipient made decisions. In the conditions FREE and COSTLY, donors had the option to request 2nd-order information about one of these three decisions, by clicking the button 'more information' below one of the recipient's decisions. The 2nd-order information that was shown after clicking the button, consisted of 1st-order information that the current recipient had when he made this decision (again, see Supplementary Information, section S4 for screenshots, available on the journal's website at www.ehbonline.org). In condition FREE, 2nd-order information was available without a cost; in condition

COSTLY, the donor was charged 5 points to obtain this information. After they decided whether to request and observe 2nd-order information, donors made their decision whether or not to help (GIVE or DENY). In early rounds of the game, when three pieces of information were not yet available, missing data were displayed as a hyphen.

2.2. Statistical analyses

The results presented in Table 1 are based on three logistic generalized linear mixed models with 'subject nested in group' as random factor. Model 1 was fitted to only those decisions in which 2nd-order information was unavailable (*i.e.* in the CONTROL condition and in cases where no such information had been requested). Model 2 only considers decisions for which donors had requested 2nd-order information about one of their recipients' DENY decisions. Similarly, Model 3 only considers those decisions for which donors had requested 2nd-order information about one of their recipients' GIVE decisions. In these models, 1st- and 2nd-order information reflected the fraction of GIVE decisions in the three decisions displayed to the donor, and the donor's own recent behavior reflected the fraction of GIVE decisions in the donor's two most recent ones. The current balance in the donor's account was divided by 3000 points (the initial endowment) before it was entered in the regression.

The classification procedure underlying Figs. 2 and 3 considers the decisions of each of the subjects separately. First, we tested whether the content of 1st-order information (fraction of recipient's GIVE decisions in his most recent three decisions) or the donor's own recent behavior (fraction of donor's GIVE decisions in her most recent two decisions) had a significant ($P < 0.05$) effect on helping behavior. To this end, we fitted a logistic generalized linear model to the decisions to help, including – apart from the content of 1st-order information about the recipient and the donor's own recent behavior – 'round' to control for changes in cooperation rates over the course of the game. We recorded whether this model detected a (positive) effect of 1st-order information or a (negative) effect of the donor's own recent behavior. Subjects were classified as 'first-order conditional cooperators' or 'cautious defectors', accordingly. Second, to identify 'second-order conditional cooperators', we tested whether 2nd-order information significantly affected donor's decisions to help. We separately considered decisions following requests for 2nd-order information about GIVE or DENY decisions. When a subject requested either of these types of information at least five times, we tested whether it affected their behavior by fitting a Bayesian generalized model to these decisions, using the content of 2nd-order information (*i.e.* the helpfulness of the recipients' recipients) as the only predictor. We recorded whether these models detected a positive (after a request for GIVE) or a negative (after a request for DENY) effect on helping rates. We used Bayesian regression to avoid issues of linear separation in our data. Fisher-exact tests comparing frequencies of help in case of negative (0 or 1 times help) or positive (2 or 3 times help) 2nd-order information led to very similar results in detecting whether this information had a significant effect on an individual's decisions to help. Finally, where the above procedures detected no significant effects yet, we classified subjects as 'unconditional cooperators' or 'unconditional defectors' if their overall helping rates were higher than 90% or lower than 10%, respectively.

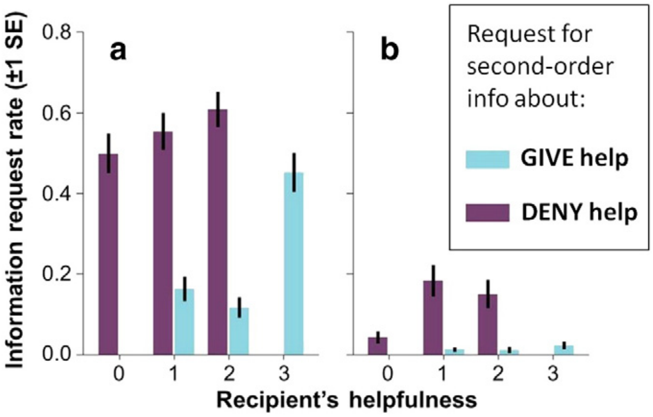

Fig. 1. Donors mostly seek 2nd-order information about defection decisions. Graphs show the frequency of donors' requests for 2nd-order information when this was (**a**) free or (**b**) costly. Blue and purple bars indicate average individual request rates (± 1 SE) for GIVE and DENY decisions, respectively. Recipient's helpfulness on the horizontal axis reflects the number of GIVE decisions out of the three decisions displayed to the donor. Request rates were higher when information was free, but in both conditions donors request more 2nd-order information about DENY decisions (compare the blue and purple bars), and more information about recipients with intermediate helpfulness scores.

3. Results

Overall helping rates were 55%, which is comparable to helping rates in studies on indirect reciprocity using similar parameters (Ule et al., 2009). Average helping rates in groups were slightly higher when 2nd-order information was present (CONTROL: 0.476; FREE: 0.600;

Table 1
Determinants of reputation-based cooperation.

	Model		
	1	2	3
First-order information	4.439 ***	5.422 ***	2.090 ***
Second-order information about a DENY decision		−2.798 ***	
Second-order information about a GIVE decision			1.017 *
Recent behavior of donor	0.825 ***	0.241	1.882 ***
FREE condition	−0.106		
COSTLY condition	−0.083	0.879 **	0.125
Round	−1.465 ***	−0.454	−2.272 ***
Previously received help as recipient	0.632 ***	0.418 *	0.381
Current total earnings	0.632 ***	0.411 *	0.269
Intercept	−3.280 ***	0.154	−1.271
N	5745	1304	649

Each of the columns presents estimates of a logistic generalized linear mixed model fit to decisions to help, with 'subject nested in group' as random effect (see Material and Methods for details). Model (1) shows overall responses to 1st-order information and own recent behavior, when no 2nd-order information was requested or available. Models (2) and (3) consider only those decisions in which donors had requested 2nd-order information about a recipient's DENY and GIVE decision, respectively. In these two models, the FREE condition was used as the baseline. In each of the models 1–3, we control for the round number and the donors' current total earnings and most recent experience as a recipient. The positive effect of the donors' own recent helping behavior suggests consistency in behavior throughout the experiment as subjects that recently helped are more likely to help again.

Significance codes:
* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

COSTLY: 0.513), but not significantly so (t-test: $P > 0.135$). These rates tended to decrease over the rounds of the game when 2nd-order information was costly or unavailable (logistic generalized linear model: $P < 0.001$). In contrast, with free 2nd-order information the average helping rates did not decrease except in the final rounds, presumably due to end effects (see Supplementary Fig. 1 for the dynamics of helping and statistical details, available on the journal's website at www.ehbonline.org).

Donors frequently requested 2nd-order information before making their helping decisions (GIVE/DENY), albeit less so when 2nd-order information was costly (Fig. 1). Overall, requests were strongly biased towards 2nd-order information about defection decisions: for respectively 62.5% (FREE) and 87.5% (COSTLY) of all requests, donors consulted information that their recipients had when making a DENY decision (as opposed to a GIVE decision; χ^2 test: $P < 0.001$). Moreover, 2nd-order information was most frequently requested when 1st-order information showed intermediate levels of helpfulness (logistic generalized linear model: $P < 0.001$, see Supplementary Information, section 2 for details, available on the journal's website at www.ehbonline.org). This suggests that 2nd-order information serves as a 'tie-breaker' when 1st-order information is not decisive.

Table 1 summarizes the aggregate effects of 1st- and 2nd-order information about helping decisions. Subjects reacted to 1st-order information in a clearly reciprocal way: donors channeled help towards recipients that had frequently helped in the past, and away from recipients that rarely helped (Table 1, column 1). When 2nd-order information was requested about a DENY decision, its content had a strong effect on helping behavior. Helping rates were high when a DENY decision was aimed at a defector, and substantially decreased when aimed at a cooperator (Table 1, column 2; Supplementary Fig. 2, available on the journal's website at www.ehbonline.org). When 2nd-order information was requested about a GIVE decision, the content of this information had a less pronounced yet significant effect: helping rates were higher when recipients helped a more cooperative individual (Table 1, column 3; Supplementary Fig. 3, available on the journal's website at www.ehbonline.org). Together, the marked aggregate responses to 1st- and 2nd-order information provide strong support for the mechanisms that can explain stable reputation-based cooperation. Help is directed towards those who denied help to defectors or helped cooperative individuals.

Do the aggregate effects we observe here reflect a commonly held strategy to assess and respond to reputations, or do individuals differ with respect to their strategies? To address this issue we analyze the helping decisions of each of the 160 subjects separately. We test how a donor's helping decisions depend on 1st- and 2nd-order information about his recipients and on the donor's own recent behavior (reflecting his current reputation). First we test whether a donor's tendency to help increased with his recipients' helpfulness or decreased with the donor's own recent helpfulness, identifying 'first-order conditional cooperators' and 'cautious defectors'. Second, if a donor requested information on at least five occasions we fit a separate model to his resulting decisions to test for his sensitivity to this information, identifying 'second-order conditional cooperators' (see Materials and Methods for details). Finally, if for a subject we detect no significant effects yet, we check if his overall helping rate was higher than 90% or lower than 10% to identify 'unconditional cooperators' or 'unconditional defectors', respectively. This three-step statistical procedure categorized 131 out of 160 participants (81.9%).

Strategies from almost all possible categories were used by the subjects, for each experimental condition (Fig. 2). The most frequent strategies are first-order conditional cooperators that help those recipients who helped others in the past (Fig. 2, segment *i*). However, when 2nd-order information was available, a substantial proportion of individuals are classified as second-order conditional cooperators, actively seeking 2nd-order information and significantly responding to it (Fig. 2, segments *iv*, *v* and *vi*). All but one of these individuals responded to 2nd-order information about DENY decisions, rewarding individuals who denied help to defectors (a key element of 'standing' strategies). One individual significantly responded to 2nd-order information about GIVE decisions, rewarding individuals who gave help to cooperators (a key element of 'judging' strategies). We also observed strategies that condition behavior on their own recent decisions (Fig. 2, segments *ii* and *iii*), helping only in order to maintain their own reputation. These 'cautious defectors' are, just like unconditional cooperators and unconditional defectors (Fig. 2, segments *C* and *D*) relatively rare.

Fig. 3 illustrates the average responses to 1st- and 2nd-order information of three identified strategies. First-order conditional cooperators respond strongly positively to the recent helping behavior of their recipients, with helping rates rising from 0.04 when a recipient made three DENY decisions, to 0.81 when a recipient made three GIVE decisions (Fig. 3A, blue squares). Cautious defectors decreased their helping rates from 0.60 to 0.15 after giving help twice (Fig. 3B, green triangles). Second-order conditional cooperators who considered the information about DENY decisions showed a strong response to the contents of 2nd-order information (Fig. 3C). In particular, 'justified defection' tended to be rewarded: helping rates were high when the recipient's DENY decision was aimed at a defector and low when the recipient's DENY decision was aimed at a cooperator (helping rates were intermediate when no 2nd-order information was requested).

4. Discussion

Our results can be summarized in three main points. First, we show that many people actively seek 2nd-order information about past behavior of other people when deciding whether to give them help. Some people seek this information even when this is costly. This interest in 2nd-order information seems indicative of an interest in the motivation behind someone's decision; what information did someone have to

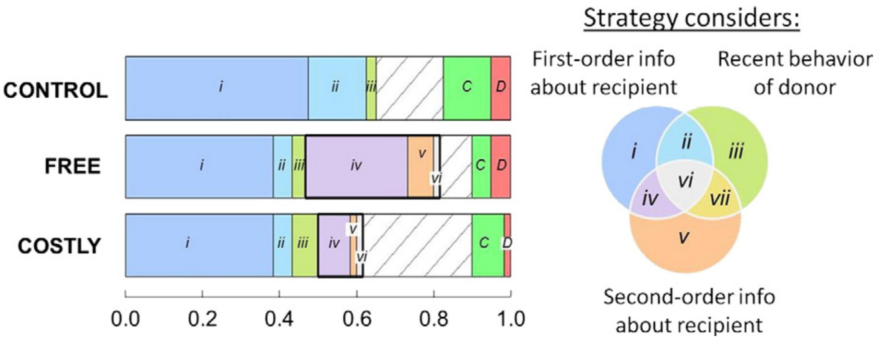

Fig. 2. Diversity of strategies in reputation-based cooperation. Bars show the prevalence of the various strategies that may condition cooperation on 1st- and 2nd-order information about the recipient's past behavior, and on the donor's own behavior. Segments of the Venn diagram with roman numerals characterize the strategies with respect to what affected their helping decisions: (1) 1st-order (blue) and (2) 2nd-order information (orange) about the recipient, and (3) donor's own recent helping (green). In the bars, thick black boxes highlight the strategies using 2nd-order information, hatched areas refer to uncategorized subjects, and the letters *C* and *D* respectively reflect 'unconditional cooperators' and 'unconditional defectors'.

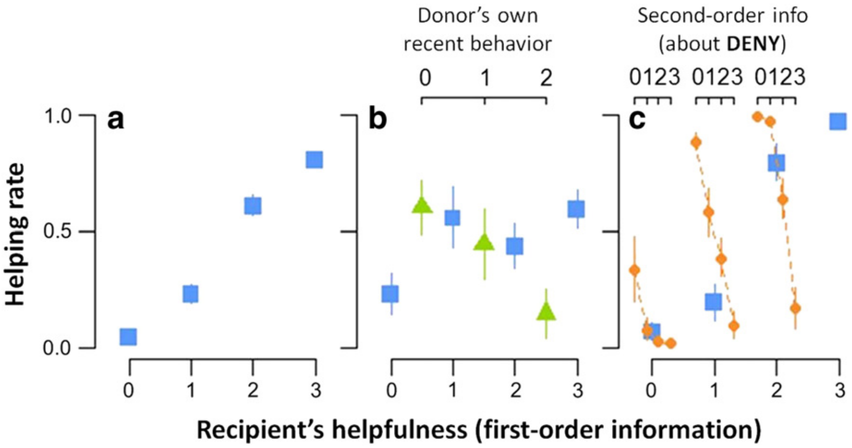

Fig. 3. Helping behavior of three prominent strategies identified by our classification procedure (see Supplementary Fig. 4 for all common strategies *i*–*iv*, *C* and *D*, available on the journal's website at www.ehbonline.org). In each of the panels, blue squares show mean helping rates of individuals as a function of the helpfulness of their recipients when no 2nd-order information was requested. Panel (**a**) shows the helping rates of 'first-order conditional cooperators' who condition their cooperation solely on the previous behavior of their recipients (strategy *i* of Fig. 2). Panel (**b**) shows the helping rates of 'cautious defectors' who only respond to their own recent behavior (green triangles; strategy *iii* of Fig. 2) and help more often when their own reputation becomes bad. Panel (**c**) shows the helping rates of 'second-order conditional cooperators' who consider and react to the information about DENY decisions (strategies *iv*–*vi* of Fig. 2); connected orange dots indicate how mean helping rates vary with the observed helpfulness of the recipients' recipients (note that information about DENY decisions is not available for recipients with three GIVE decisions). The sharply declining trends indicate that helping strongly depended on whether a DENY decision was aimed at a defector (2nd-order information = 0) or a cooperator (2nd-order information = 3). For each of the dots in each panel, error bars indicate ±1 SE.

make their decision and how did they act on that information? Second, people are mostly interested in 2nd-order information about defection decisions. This information strongly affects helping behavior, and justified defection is generally rewarded. Third, our analysis reveals strong individual variation in behavioral strategies. Some individuals condition their cooperation solely on the 1st-order information about past behavior of their recipients, while others also consider the 2nd-order information about the context of this behavior. Together, these findings provide strong empirical support for mechanisms proposed in theoretical models explaining stable human cooperation through indirect reciprocity.

Our results indicate that 'standing' strategies may play a more important role in human reputation-based cooperation than 'judging' strategies. Our subjects are more interested in motivations behind defection and these have a larger effect on cooperative decision making (Table 1; Supplementary Figs. 2 and 3, available on the journal's website at www.ehbonline.org). Moreover, it is conceivable that strategic variation of the kind that we observe can have marked consequences on the emergence and stability of reputation-based cooperation. In order to identify evolutionarily stable strategies the theoretical models of reputation-based cooperation typically assumed that all members of a population evaluate reputations in the same way (e.g., Nowak & Sigmund, 1998; Ohtsuki & Iwasa, 2007; Panchanathan & Boyd, 2004). Our study, however, highlights the possibility that cooperation dynamics depend on the specific mix of strategies in a group of individuals. Our results call for new theoretical work extending initial explorations of how individual variation in reputation assessment within a population affects the dynamics of reciprocal helping (Uchida & Sigmund, 2010) and assessing how mixtures of different strategies could affect the emergence and stability of cooperation. One obvious consideration is how this stability is affected when individuals can actively choose with whom to interact (Rand, Arbesman, & Christakis, 2011). If individuals could gauge the strategies of others and bias their interactions accordingly, clusters of individuals using motivation-based (2nd-order) strategies may achieve stable cooperation, whereas cooperation among behavior-based (1st-order) strategies may break down in the presence of defectors.

Although the relatively rich information conditions of our experiment allow for complex strategies, we consider a simplification of the concept of a 'reputation' suggested by evolutionary models. In reality, reputation builds up over various contexts of interaction (Macfarlan & Lyle, 2015), where factors like information reliability may play a role. The mechanisms of information transfer are crucial: one may directly observe interactions between individuals, but reputations also propagate indirectly through gossip (Sommerfeld, Krambeck, & Milinski, 2008, Sommerfeld, Krambeck, Semmann, & Milinski, 2007). Gossip is prone to error, lies, and strategic manipulation. As 2nd-order information stems from interactions in the more distant past and between more socially distant individuals, it could be less reliable than 1st-order information (Ohtsuki & Iwasa, 2004; Panchanathan, 2011). Strategies based on 2nd-order information might therefore be more strongly affected by inaccurate reputational information, undermining the efficiency of these strategies in supporting reputation-based cooperation.

Individuals interacting in small groups often base their decisions not only on reputations, but also on previous personal encounters with others. Indeed, cooperative behavior may then depend on the interplay between direct and indirect reciprocity (Molleman et al., 2013; Roberts, 2008). From this perspective, our understanding of human cooperation through reciprocal helping could benefit from more general insights in how reputational information is weighted with reliability and integrated with information from direct experience.

Competing interests

We have no competing interests.

Author contributions

V.S., L.M., A.U. and M.E. designed the experiments, V.S., L.M. and M.E. carried out the experiments and V.S., L.M., A.U. and M.E. wrote the paper.

Acknowledgments

We are grateful to Franjo Weissing for helpful feedback on earlier drafts of this manuscript. We also thank Benjamin Beranek, Simon Gächter, Felix Kölle, Liam Pollock, Jonathan Schulz, Till Weber and Ori Weisel for useful discussions and comments.

References

Alexander, R. D. (1987). *The biology of moral systems.* Transaction Publishers.

Bolton, G. E., Katok, E., & Ockenfels, A. (2005). Cooperation among strangers with limited information about reputation. *Journal of Public Economics*, *89*, 1457–1468. http://dx.doi.org/10.1016/j.jpubeco.2004.03.008.

Brandt, H., & Sigmund, K. (2004). The logic of reprobation: Assessment and action rules for indirect reciprocation. *Journal of Theoretical Biology*, *231*, 475–486. http://dx.doi.org/10.1016/j.jtbi.2004.06.032.

Brandt, H., & Sigmund, K. (2005). Indirect reciprocity, image scoring, and moral hazard. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 2666–2670. http://dx.doi.org/10.1073/pnas.0407370102.

Clutton-Brock, T. (2009). Cooperation between non-kin in animal societies. *Nature*, *462*, 51–57. http://dx.doi.org/10.1038/nature08366.

Darwin, C. (1871). *The descent of man.* London: John Murray.

dos Santos, M., Braithwaite, V. A., & Wedekind, C. (2014). Exposure to superfluous information reduces cooperation and increases antisocial punishment in reputation-based interactions. *Frontiers in Ecology and Evolution*, *2*, 41.

Dugatkin, L. A. (1997). *Cooperation among animals.* Oxford: Oxford University Press.

Engelmann, D., & Fischbacher, U. (2009). Indirect reciprocity and strategic reputation building in an experimental helping game. *Games and Economic Behavior*, *67*, 399–407. http://dx.doi.org/10.1016/j.geb.2008.12.006.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*, 171–178. http://dx.doi.org/10.1007/s10683-006-9159-4.

Fischbacher, U., & Gachter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review*, *100*, 541–556. http://dx.doi.org/10.1257/aer.100.1.541.

Gavrilets, S. (2015). Collective action problem in heterogeneous groups. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *370*, 20150016. http://dx.doi.org/10.1098/rstb.2015.0016.

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology*, *206*, 169–179. http://dx.doi.org/10.1006/jtbi.2000.2111.

Hamilton, W. (1964). Genetical evolution of social behaviour 2. *Journal of Theoretical Biology*, *7*, 17. http://dx.doi.org/10.1016/0022-5193(64)90039-6.

Hartig, B., Irlenbusch, B., & Koelle, F. (2015). Conditioning on what? Heterogeneous contributions and conditional cooperation. *Journal of Behavioral and Experimental Economics*, *55*, 48–64. http://dx.doi.org/10.1016/j.socec.2015.01.001.

Kurzban, R., & Houser, D. (2005). Experiments investigating cooperative types in humans: A complement to evolutionary theory and simulations. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 1803–1807.

Leimar, O., & Hammerstein, P. (2001). Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society of London B: Biological Sciences*, *268*, 745–753.

Macfarlan, S. J., & Lyle, H. F. (2015). Multiple reputation domains and cooperative behaviour in two Latin American communities. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *370*, 20150009. http://dx.doi.org/10.1098/rstb.2015.0009.

Mashima, R., & Takahashi, N. (2005). Is the enemy's friend an enemy too? Theoretical and empirical approach toward the effect of second-order information on indirect reciprocity. *Sociological Theory and Methods*, *20*, 177–195.

McNamara, J. M., & Leimar, O. (2010). Variation and the response to variation as a basis for successful cooperation. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *365*, 2627–2633. http://dx.doi.org/10.1098/rstb.2010.0159.

Milinski, M., Semmann, D., Bakker, T. C. M., & Krambeck, H. J. (2001). Cooperation through indirect reciprocity: Image scoring or standing strategy? *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *268*, 2495–2501. http://dx.doi.org/10.1098/rspb.2001.1809.

Molleman, L., van den Berg, P., & Weissing, F. J. (2014). Consistent individual differences in human social learning strategies. *Nature Communications*, *5*, 3570. http://dx.doi.org/10.1038/ncomms4570.

Molleman, L., van den Broek, E., & Egas, M. (2013). Personal experience and reputation interact in human decisions to help reciprocally. *Proceedings of the Royal Society B: Biological Sciences*, *280*, 20123044. http://dx.doi.org/10.1098/rspb.2012.3044.

Noe, R., & Hammerstein, P. (1994). Biological markets — Supply-and-demand determine the effect of partner choice in cooperation, mutualism and mating. *Behavioral Ecology and Sociobiology*, *35*, 1–11.

Nowak, M. A., & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, *393*, 573–577. http://dx.doi.org/10.1038/31225.

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, *437*, 1291–1298. http://dx.doi.org/10.1038/nature04131.

Ohtsuki, H., & Iwasa, Y. (2004). How should we define goodness? Reputation dynamics in indirect reciprocity. *Journal of Theoretical Biology*, *231*, 107–120.

Ohtsuki, H., & Iwasa, Y. (2006). The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology*, *239*, 435–444.

Ohtsuki, H., & Iwasa, Y. (2007). Global analyses of evolutionary dynamics and exhaustive search for social norms that maintain cooperation by reputation. *Journal of Theoretical Biology*, *244*, 518–531. http://dx.doi.org/10.1016/j.jtbi.2006.08.018.

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a sword — Self-governance is possible. *The American Political Science Review*, *86*, 404–417. http://dx.doi.org/10.2307/1964229.

Pacheco, J. M., Santos, F. C., & Chalub, F. A. C. (2006). Stern-judging: A simple, successful norm which promotes cooperation under indirect reciprocity. *PLoS Computational Biology*, *2*, e178.

Panchanathan, K. (2011). Two wrongs don't make a right: The initial viability of different assessment rules in the evolution of indirect reciprocity. *Journal of Theoretical Biology*, *277*, 48–54.

Panchanathan, K., & Boyd, R. (2003). A tale of two defectors: The importance of standing for evolution of indirect reciprocity. *Journal of Theoretical Biology*, *224*, 115–126.

Panchanathan, K., & Boyd, R. (2004). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature*, *432*, 499–502. http://dx.doi.org/10.1038/nature02978.

Rand, D. G., Arbesman, S., & Christakis, N. A. (2011). Dynamic social networks promote cooperation in experiments with humans. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 19193–19198. http://dx.doi.org/10.1073/pnas.1108243108.

Roberts, G. (2008). Evolution of direct and indirect reciprocity. *Proceedings of the Royal Society B: Biological Sciences*, *275*, 173–179. http://dx.doi.org/10.1098/rspb.2007.1134.

Seinen, I., & Schram, A. (2006). Social status and group norms: Indirect reciprocity in a repeated helping experiment. *European Economic Review*, *50*, 581–602. http://dx.doi.org/10.1016/j.euroecorev.2004.10.005.

Sigmund, K. (2012). Moral assessment in indirect reciprocity. *Journal of Theoretical Biology*, *299*, 25–30. http://dx.doi.org/10.1016/j.jtbi.2011.03.024.

Maynard Smith, J. (1964). Group selection and kin selection. *Nature*, *201*, 1145–1147. http://dx.doi.org/10.1038/2011145a0.

Sommerfeld, R. D., Krambeck, H. -J., & Milinski, M. (2008). Multiple gossip statements and their effect on reputation and trustworthiness. *Proceedings of the Royal Society B: Biological Sciences*, *275*, 2529–2536. http://dx.doi.org/10.1098/rspb.2008.0762.

Sommerfeld, R. D., Krambeck, H. -J., Semmann, D., & Milinski, M. (2007). Gossip as an alternative for direct observation in games of indirect reciprocity. *Proceedings of the National Academy of Sciences*, *104*, 17435–17440.

Sugden, R. (1986). *The economics of rights, co-operation and welfare.* Oxford: Basil Blackell.

Takahashi, N., & Mashima, R. (2006). The importance of subjectivity in perceptual errors on the emergence of indirect reciprocity. *Journal of Theoretical Biology*, *243*, 418–436. http://dx.doi.org/10.1016/j.jtbi.2006.05.014.

Trivers, R. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, *46*, 35–57.

Uchida, S., & Sigmund, K. (2010). The competition of assessment rules for indirect reciprocity. *Journal of Theoretical Biology*, *263*(1), 13–19.

Ule, A. (2008). *Partner choice and cooperation in networks: Theory and experimental evidence.* Berlin, Heidelberg: Springer.

Ule, A., Schram, A., Riedl, A., & Cason, T. N. (2009). Indirect punishment and generosity toward strangers. *Science*, *326*, 1701–1704. http://dx.doi.org/10.1126/science.1178883.

van den Berg, P., Molleman, L., Junikka, J., Puurtinen, M., & Weissing, F. J. (2015a). Human cooperation in groups: Variation begets variation. *Science Reports*, *5*, 16144. http://dx.doi.org/10.1038/srep16144.

van den Berg, P., Molleman, L., & Weissing, F. J. (2015b). Focus on the success of others leads to selfish behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *112*, 2912–2917. http://dx.doi.org/10.1073/pnas.1417203112.

Wedekind, C., & Braithwaite, V. A. (2002). The long-term benefits of human generosity in indirect reciprocity. *Current Biology*, *12*, 1012–1015. http://dx.doi.org/10.1016/S0960-9822(02)00890-4.

Wedekind, C., & Milinski, M. (2000). Cooperation through image scoring in humans. *Science*, *288*, 850–852. http://dx.doi.org/10.1126/science.288.5467.850.

Wolf, M., van Doorn, G. S., & Weissing, F. J. (2008). Evolutionary emergence of responsive and unresponsive personalities. *Proceedings of the National Academy of Sciences*, *105*, 15825–15830. http://dx.doi.org/10.1073/pnas.0805473105.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personal and Social Psychology*, *51*, 110–116. http://dx.doi.org/10.1037//0022-3514.51.1.110.