PROCEEDINGS B

rspb.royalsocietypublishing.org

Research

Cite this article: Tanaka H, Ohtsuki H, Ohtsubo Y. 2016 The price of being seen to be just: an intention signalling strategy for indirect reciprocity. *Proc. R. Soc. B* **283**: 20160694.
http://dx.doi.org/10.1098/rspb.2016.0694

Received: 25 March 2016
Accepted: 29 June 2016

Subject Areas:
behaviour, evolution

Keywords:
indirect reciprocity, cooperation, intention signalling, mind reading

Author for correspondence:
Yohsuke Ohtsubo
e-mail: yohtsubo@lit.kobe-u.ac.jp

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2016.0694 or via http://rspb.royalsocietypublishing.org.

THE ROYAL SOCIETY PUBLISHING

The price of being seen to be just: an intention signalling strategy for indirect reciprocity

Hiroki Tanaka[1,2], Hisashi Ohtsuki[3] and Yohsuke Ohtsubo[1]

[1]Department of Psychology, Graduate School of Humanities, Kobe University, 1-1 Rokkodai-cho, Nada, Kobe 657-8501, Japan
[2]Japan Society for the Promotion of Science, Tokyo, Japan
[3]Department of Evolutionary Studies of Biosystems, School of Advanced Sciences, SOKENDAI (The Graduate University for Advanced Studies), Shonan Village, Hayama 240-0193, Kanagawa, Japan

YO, 0000-0003-2074-0244

Cooperation among strangers is a marked characteristic of human sociality. One prominent evolutionary explanation for this form of human cooperation is indirect reciprocity, whereby each individual selectively helps people with a 'good' reputation, but not those with a 'bad' reputation. Some evolutionary analyses have underscored the importance of second-order reputation information (the reputation of a current partner's previous partner) for indirect reciprocity as it allows players to discriminate justified 'good' defectors, who selectively deny giving help to 'bad' partners, from unjustified 'bad' defectors. Nevertheless, it is not clear whether people in fact make use of second-order information in indirect reciprocity settings. As an alternative, we propose the intention signalling strategy, whereby defectors are given the option to abandon a resource as a means of expunging their 'bad' reputation. Our model deviates from traditional modelling approaches in the indirect reciprocity literature in a crucial way—we show that first-order information is sufficient to maintain cooperation if players are given an option to signal their intention. Importantly, our model is robust against invasion by both unconditionally cooperative and uncooperative strategies, a first step towards demonstrating its viability as an evolutionarily stable strategy. Furthermore, in two behavioural experiments, when participants were given the option to abandon a resource so as to mend a tarnished reputation, participants not only spontaneously began to use this option, they also interpreted others' use of this option as a signal of cooperative intent.

1. Introduction

Human beings are a highly cooperative species [1–3]. Small acts of kindness towards strangers such as giving directions to a traveller or offering one's seat to an elderly person are pervasive in human societies. It is not even unheard of for people to risk their own lives to save the life of a stranger. Such instances of altruistic behaviour in one-shot interactions cannot be explained by reciprocal altruism [4,5]. Nevertheless, if altruists selectively help other altruists, selfless acts directed towards strangers are evolvable. This system is known as indirect reciprocity—if I help you, someone else will help me [6–13]. A simplest strategy for indirect reciprocity is the image-scoring strategy, which confers a 'good' reputation upon, and preferentially cooperates with, other cooperative players [7,8]. Although this strategy might appear to endorse discriminating between cooperative and uncooperative players, it simultaneously fosters disincentives to do so [10,14]. For example, suppose that one player refrained from helping a bad player. This player will acquire a 'bad' reputation for his/her non-cooperative behaviour on the next round. Therefore, each player would be better off by cooperating with a bad player. To make matters worse, the initial defection against a bad player invites a chain of unnecessary defections, which undermines the cooperative equilibrium [12,13].

This problem can be avoided by the standing strategy, an evolutionarily stable strategy for indirect reciprocity, whereby defection against players in 'bad' standing is justified and distinguished from defection against players in 'good' standing [10–13]. In fact, an exhaustive search of evolutionarily stable strategies for indirect reciprocity has revealed that the cooperative equilibrium with the highest net pay-off can be maintained by only eight (out of 4096) strategies, and that all of them, collectively called the 'leading eight', distinguish justified from unjustified defection [12,13]. It is important to note that the distinction between these two types of defection requires second-order information (a current partner's previous partner's reputation). It thus implicitly assumes that people use second-order information to determine whether an actor withheld help with a justifiable intention or a genuinely uncooperative intention. Although it is true that people have a sophisticated capacity for theory of mind (or social intelligence) [15,16], people enrolled in experimental games do not usually engage in deep strategic reasoning [17]. In fact, empirical evidence concerning whether people readily make use of second-order information is mixed. Although earlier studies reported negative results [18,19], there are some recent studies reporting positive results [20,21]. In sum, although theoretical works conceive second-order information as key to stabilizing indirect reciprocity, empirical works do not unequivocally indicate that people readily make use of second-order information.

It is noteworthy that traditional models in the indirect reciprocity literature have implicitly assumed that people are passive subjects of evaluation (at least in terms of their choices to cooperate or defect). Quite the opposite appears to be the case, however, as people have been observed to actively manage the impressions they make upon others. For example, people behave more cooperatively in the presence of reputational benefits [14,22–27]. In Milinski and colleagues' indirect reciprocity experiment [18], justified defectors (i.e. people who withheld help from a previous defector) subsequently increased their cooperative behaviour in an apparent attempt to recover a tarnished reputation. Likewise, people tend to act in a more altruistic manner when their moral worth is damaged [28,29]. This type of reputation recovery strategy, which was modelled as contrite tit-for-tat (*CTFT*), yields a more efficient cooperative equilibrium than the standing strategy [11]. Although *CTFT* accepts a 'bad' reputation at least once, people have been shown to react to their social predicaments more immediately by offering apologies [30–33] and/or inflicting self-punishment [33–36]. Justified defectors may be inclined to use these sorts of signals to communicate their non-malicious intent. In this article, we first present an evolutionary game analysis which shows that an intention signalling strategy (*intSIG*) can stabilize cooperation by indirect reciprocity. We then report the results of two experiments showing that people actually behave in an *intSIG*-like manner.

2. Evolutionary game analysis

To model the *intSIG* strategy, the standard indirect reciprocity setting was modified as follows: a donor decides whether to incur a cost (c) to confer a benefit (b) on a recipient ($b > c > 0$). When a donor decides to withhold help, the donor subsequently decides whether to spend a resource, s ($=c$), to produce a signal. The donor produces the signal by abandoning the resource that he/she saved by withholding help. These signallers maintain their good standing even though they withheld help. By contrast, donors who opt not to use the signal after withholding help lose their good standing. The signal cost (s) is set to be equal to the cost of cooperation (c). If the cost of the signal was cheaper than that of cooperation, fake signallers could maintain their good standing by producing cheap signals. Thus, setting s equal to c curtails the incentive to fake the signal. Note that *intSIG* does not rely on second-order information because it can restore an endangered reputation immediately after the act that puts the reputation in danger. Moreover, it does not rely on the observers' cognitive ability to use second-order information. Instead, it assumes the coevolution of a signal sending propensity and receivers' signal-reading ability.

To see how *intSIG* works in repeated interactions, here we describe a simplified version of evolutionary game analyses (see the electronic supplementary material for more formal analyses). Suppose there is a population of individuals who randomly form pairs on a round-by-round basis. Further suppose that on each round the players are randomly assigned to play as either a donor or recipient. When everyone in the population uses *intSIG*, they earn either $-c$ (as a donor) or b (as a recipient) with the same probability every round, so the net pay-off is $(b-c)/2$ times $1/(1-\omega)$, where ω is the probability of having another round with any member of the population, and hence $1/(1-\omega)$ is the expected number of rounds. Consider two potential invaders: unconditional defectors (*ALLD*) who neither cooperate nor signal, and unconditional cooperators (*ALLC*) who always cooperate (and thus never signal). A rare *ALLD* player, who is initially in good standing, obtains b as long as it continues to play the recipient role from the first round; once it plays as a donor it will be in bad standing forever and never receive cooperation. It is easy to confirm that the average number of rounds that *ALLD* receives cooperation is $(1/2)+\omega(1/2)^2+\cdots=1/(2-\omega)$. Accordingly, *ALLD*'s net pay-off is $b/(2-\omega)$. Therefore, the comparison between $(b-c)/[2(1-\omega)]$ and $b/(2-\omega)$ reveals that *intSIG*'s net pay-off exceeds *ALLD*'s net pay-off if ω is sufficiently large: $\omega > 2c/(b+c)$.

By contrast, *ALLC* players obtain the same net pay-off of $(b-c)/[2(1-\omega)]$ as *intSIG*. However, if there is even a small chance of committing errors in the implementation of their cooperative intent, a pay-off difference arises between *intSIG* and *ALLC*, and *intSIG* can resist invasion by *ALLC*. Let us pay attention to this pay-off difference. After committing an implementation error as a donor, *intSIG* abandons c to produce the signal. Therefore, the pay-off of *intSIG*, given that it played as a donor and committed an error in this single round, is $-c$ (the cost of the signal). Thanks to this signal, however, *intSIG* can keep its good standing, and the error no longer casts a shadow over *intSIG*'s future pay-off consequences. By contrast, when *ALLC* commits an implementation error as a donor, *ALLC* does not signal and hence pays nothing. But it casts a shadow over *ALLC*'s future: *ALLC* is degraded to a bad standing and continues to miss the benefit b of cooperation until the next time it plays as a donor, at which point it can recover its good standing. On average, *ALLC* misses receiving the benefit of cooperation $\omega(1/2)+\omega^2(1/2)^2+\cdots=\omega/(2-\omega)$ times, resulting in the loss of $b\omega/(2-\omega)$. Therefore, the loss $b\omega/(2-\omega)$ can be greater than c if ω is sufficiently large: $\omega > 2c/(b+c)$. Interestingly, the condition that stabilizes *intSIG* against *ALLC* is identical to the condition that stabilizes *intSIG* against *ALLD*.

3. Experiments

Given the results of the evolutionary game analysis, we then tested whether people behave in an *intSIG*-like manner in two behavioural experiments. In particular, participants played the donation game [18,37] under either a signalling condition or a standing condition. In the donation game, participants were randomly paired with another putative participant on each round, and randomly assigned to the role of either donor or recipient. Donors decided whether to give their resource to the recipient. In the signalling condition, when donors decided not to give the resource for whatever reason, they were given the additional option to abandon their resource. In the standing condition, before deciding whether to give the resource, donors were presented with second-order information about the past behaviour of their recipient's previous partner. In each condition, participants had the chance to interact with recipients (i.e. pre-programed computerized partners) displaying every possible type of reputation information. Therefore, this procedure allowed us to determine the strategies that each participant employed.

(a) Hypotheses

In the signalling condition, we tested the following three hypotheses. Hypotheses 1a and 1b are about signallers' behaviours. Hypothesis 2 is about signal receivers' reaction to the signal. Hypothesis 1a is based on the operationalization of three types of defection: 'unjustified defection' (not giving the resource to a player in good standing), 'justified defection' (not giving the resource to a player in bad standing) and 'implementation error' (a computer-generated replacement of one's give choice with the not-give choice). It should be noted that *intSIG* may commit justified defection and implementation error, but not unjustified defection. Therefore, the following pattern is expected.

> Hypothesis 1a: Participants use the signal option more frequently after justified defection and implementation error than unjustified defection.

Based on the above typology of defection, defectors are categorized as two types. 'Unjustified defectors' are those who do not give the resource to recipients regardless of their standing. 'Justified defectors' are those who selectively withhold giving the resource to recipients in bad standing. However, because justified defection and unjustified defection are indistinguishable in the absence of second-order information, justified defectors need to distinguish themselves from unjustified defectors by using the signal option.

> Hypothesis 1b: Justified defectors use the signal option more frequently than unjustified defectors.

Nevertheless, to conclude that people use an *intSIG*-like strategy, we have to confirm that they also use other players' signals to discriminate good players from bad players.

> Hypothesis 2: Participants give their resources more frequently to players who, in the previous round, gave their resource (givers) or did not give it but used the signal option (signalling non-givers) than to players who defected without using the signal option (non-signalling non-givers).

In the standing condition, participants in the donor role were provided with second-order information, enabling the distinction of four types of recipients: GG, GN, NG and NN, where the left side G/N represents the past behaviour of the donor's current recipient ('gave' or 'did not give'), and the right side G/N represents the behaviour of the recipient's previous recipient ('gave' or 'did not give'). If participants distinguish justified defection from unjustified defection based on second-order information, they should discriminate NN-recipients as justified defectors from NG-recipients. Therefore, participants should give resources to GG-, GN- and NN-recipients more frequently than NG-recipients. However, if participants do not use second-order information, it is expected that participants will give resources to GG- and GN-recipients more frequently than NG- and NN-recipients.

(b) Method common to experiments 1 and 2

Participants were 107 undergraduates (62 males and 45 females) and 102 undergraduates (48 males and 54 females) in experiments 1 and 2, respectively, at a large university in Japan. There was no overlap in these two groups of participants. Three participants were omitted from each experiment because they suspected the absence of other players or did not understand the rules of the donation game. In both experiments, participants were randomly assigned to either the signalling or standing condition.

All participants first played 50 rounds of the standard donation game. This served as a practice session. In this session, participants were informed that they would take part in an experimental game with five other participants. In fact, they played the game with a computer program. In each round, participants were randomly assigned to either the donor or recipient role. When assigned to the donor role, participants received 5 Japanese yen (5 JPY $\approx$ £0.03) as an endowment, and decided whether to 'give' or 'not give' it to the current recipient. The recipients would receive 10 JPY if their donor chose 'give', but received 0 JPY if their donor chose 'not give'. Participants in the donor role received feedback regarding their current decision (e.g. 'you chose "give"') immediately after they made the decision. To introduce a small amount of implementation error, donors' give choices were replaced with not-give choices by the computer program with a small probability. If errors occurred, participants in the donor role were made immediately aware. Without the participants' knowledge, the probability of implementation error was set to 10%. In this practice session, donors only received first-order information: how their recipient behaved the last time he/she was assigned to the donor role (the data in this practice session are reported in the electronic supplementary material).

After the above practice session, participants played 100 rounds of the donation game under either the signalling or standing condition. In the signalling condition, participants were informed that they would have an extra behavioural option after choosing 'not give'. They were allowed to abandon the 5 JPY that they saved in that round. Therefore, the following first-order information about recipients' previous behaviour was made available to current donors: 'gave', 'did not give + abandoned' or 'did not give + did not abandon'. All participants were paired with recipients with 'gave', 'did not give + abandoned' and 'did not give + did not abandon' histories approximately 25, 13 and 12 times, respectively.

In the standing condition, participants were informed that they would receive additional information regarding their current recipient's previous partner's behaviour

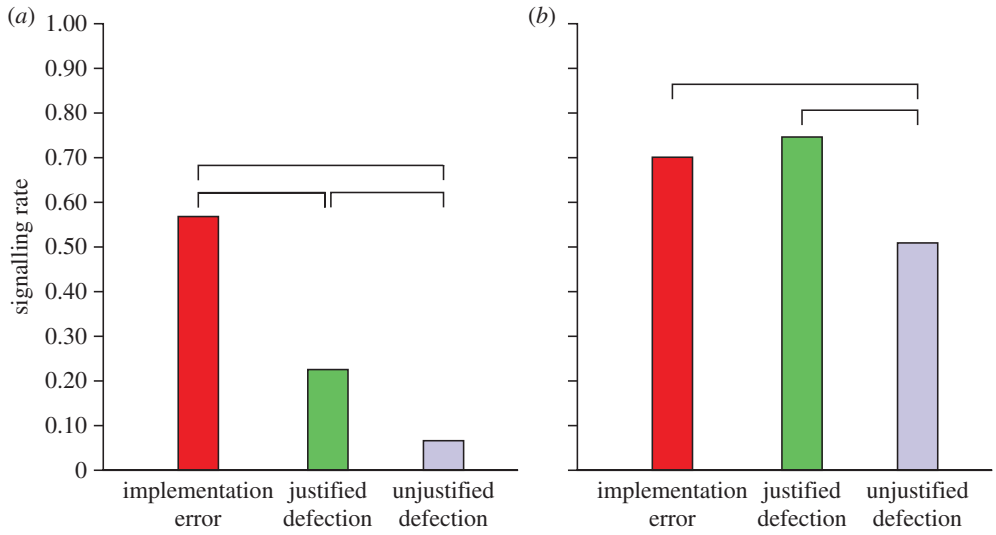

Figure 1. Relative frequency of signal use after implementation error, justified defection and unjustified defection in (*a*) experiment 1 and (*b*) experiment 2. (Online version in colour.)

(i.e. second-order information). Therefore, donors in the standing condition were informed whether the recipient 'gave' or 'did not give' a resource to their previous recipient who either 'had given' or 'had not given' a resource. The combination of these two pieces of information yielded four types of recipients: GG, GN, NG and NN. All participants were paired with GG-, GN-, NG- and NN-recipients approximately 13, 13, 12 and 12 times, respectively.

After the experimental game, we asked participants to fill out the post-experiment questionnaire to assess the strategy that each participant used (see the electronic supplementary material for details). After the post-experiment questionnaire, participants were debriefed and paid 1500 JPY.

(c) Differences between experiments 1 and 2

The two sets of experiments differed in terms of the information provided about other players' strategies. In experiment 1, participants in the recipient role were not informed about whether their donors chose to give or not give a resource to them. They were, however, informed of their own cumulative earnings after every five rounds of playing the recipient role. To further obscure the other players' strategies, cumulative earnings for the five rounds were randomly determined from a uniform distribution of 0 JPY (receiving 10 JPY from no donors) to 50 JPY (receiving 10 JPY from all five donors) with an increment of 10 JPY. Therefore, if participants in experiment 1 behave in an *intSIG*-like manner, this cannot be attributed to social learning. This suggests that *intSIG* is in participants' natural behavioural repertoire. In experiment 2, however, participants were informed of their donor's choice every round they were assigned to the recipient role. Four bogus players used either *intSIG* or the standing strategy (according to the condition), and one player used *ALLD*. Participants in experiment 2 were also made aware of their cumulative earnings throughout the game. Thus, experiment 2 tested whether the cues of other players' use of *intSIG* would enhance participants' use of *intSIG*.

(d) Results of experiment 1

We first examined whether participants used the signal option. Out of 52 participants in the signalling condition, 48 participants used the signal option at least once (45, 31 and 22 participants used it at least once after implementation error, justified defection and unjustified defection, respectively). To obtain the signalling rate, the number of signal uses by each participant was summed for each type of defection, and then divided by the total number of each type of defection that participants committed. As shown in figure 1*a*, participants used the signal option more frequently after implementation error than justified defection and unjustified defection. More importantly, participants used the signal option more frequently after justified defection than unjustified defection. These differences were significant by Fisher's exact test with the Bonferroni correction (every $p < 0.017$). Therefore, hypothesis 1a was supported.

To test hypothesis 1b, we operationally defined 'unjustified defectors' and 'justified defectors' as follows. Unjustified defectors ($n = 13$) were those who committed defection more than 80% of the time when they were paired with a partner in good standing. Among the remaining participants, justified defectors ($n = 19$) were those who committed defection more than 80% of the time when paired with a partner in bad standing. Consistent with hypothesis 1b, justified defectors used the signal option significantly more often (0.341, s.d. = 0.101) than unjustified defectors (0.007, s.d. = 0.000): $t_{30} = 3.77$, $p < 0.001$.

We then examined how participants in the donor role responded to their recipients. Recipients were categorized as 'giver', 'signalling non-giver' and 'non-signalling non-giver' based on their previous behaviour as a donor. For each participant, the giving rate to these three types of recipients was computed separately. The mean giving rate as a function of recipient type is shown in figure 2*a*. As expected, the main effect of partner type was significant ($F_{2,102} = 34.83$, $p < 0.001$), and a post hoc test by Ryan's method indicated that participants gave the resource to givers and signalling non-givers more frequently than non-signalling non-givers. In addition, participants gave the resource to givers more frequently than signalling non-givers. Therefore, hypothesis 2 was supported.

In the standing condition, participants did not distinguish justified defection from unjustified defection. Although the effect of recipient type was significant ($F_{3,153} = 20.49$, $p < 0.001$), a post hoc test by Ryan's method indicated that

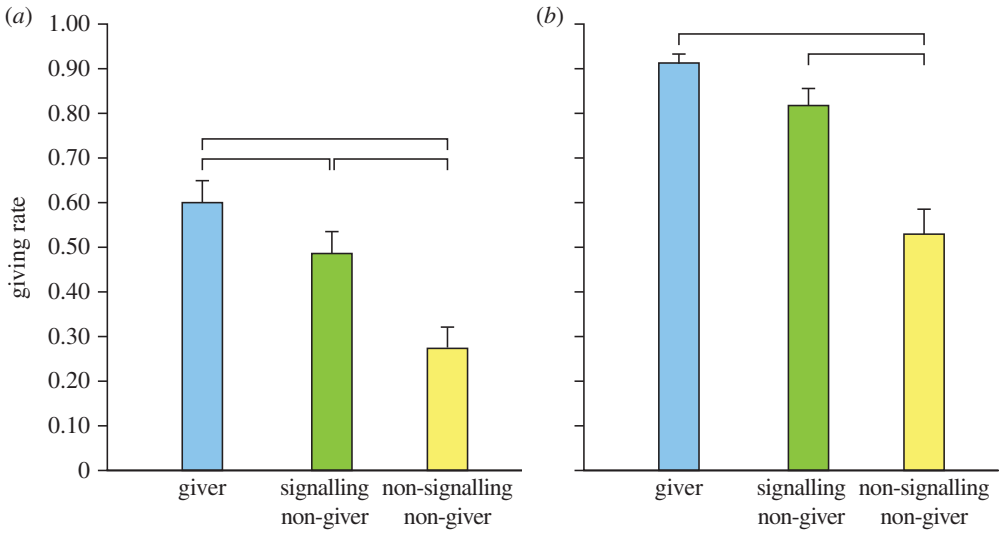

Figure 2. Mean giving rate to giver, signalling non-giver and non-signalling non-giver in the signalling condition of (*a*) experiment 1 and (*b*) experiment 2. (Online version in colour.)

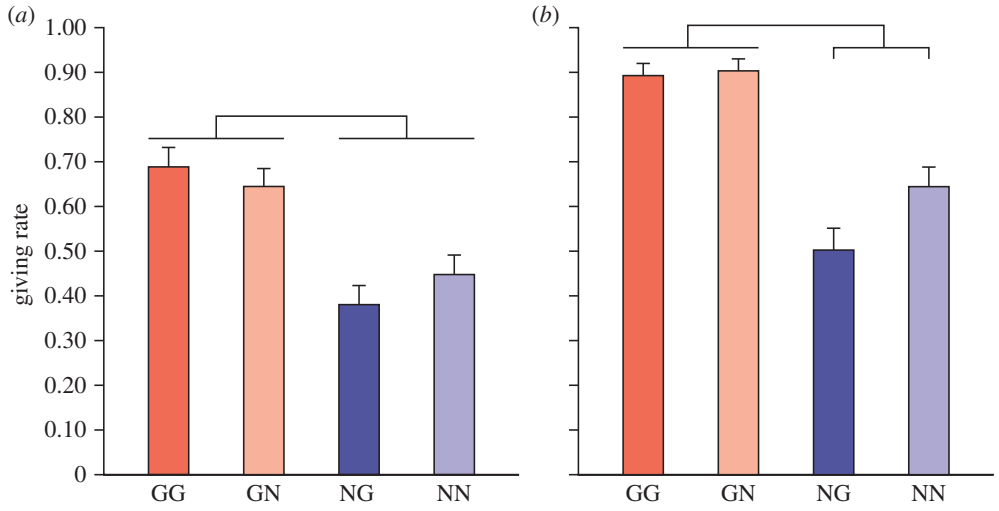

Figure 3. Mean giving rate to GG, GN, NG and NN recipient in the standing condition of (*a*) experiment 1 and (*b*) experiment 2. (Online version in colour.)

participants gave the resource to GG- and GN-recipients more frequently than NG- and NN-recipients (figure 3*a*).

(e) Results of experiment 2

In experiment 2, participants received immediate feedback about their donor's behaviour when they were assigned to the recipient role. This gave the participants a chance to infer other players' strategies. Participants' signal use increased as a result of this procedural change (figure 1*b*). However, more importantly, participants used the signal option after implementation error and justified defection significantly more often than after unjustified defection, as revealed by a Fisher's exact test with the Bonferroni correction ($p < 0.001$ for each comparison). On the other hand, signalling frequency after implementation error and justified defection did not differ ($p = 0.293$). In experiment 2, there were no participants who committed unjustified defection more than 80% of the time. Therefore, we were unable to test hypothesis 1b. Even when we relaxed the criterion of defectors to 'unjustified defection 50% of the time', only five participants were identified as unjustified defectors. These five unjustified defectors (0.48, s.d. = 0.17) used the signal option less frequently than the 18 justified defectors (0.82, s.d. = 0.08): $t_{21} = 2.14$, $p = 0.044$. Although this was not a stringent test, the result is consistent with hypothesis 1b.

In experiment 2, participants were again more likely to give the resource to givers and signalling non-givers than non-signalling non-givers (figure 2*b*). The effect of recipient type was significant ($F_{2,96} = 31.38$, $p < 0.001$), and a post hoc test indicated that participants treated givers and signalling non-givers significantly more favourably than non-signalling non-givers. Hypothesis 2 was again supported. On the other hand, in the standing condition, participants used a standing-like rule to decide whether to give the resource. Although they still favoured the GG- and GN-recipients more than the NG- and NN-recipients, they gave the resource more often to the NN-recipient (justified defector) than the NG-recipient (unjustified defector).

4. Discussion

The results of the two experiments clearly demonstrate that people are willing to manage their reputation in a costly manner as long as they are allowed to do so. This tendency

was accentuated by giving participants the chance to get acquainted with other players' strategies. Possibly, having a window into other players' strategies made the prospect of evaluation by other players more salient. In addition, participants treated signalling non-givers more favourably than non-signalling non-givers. Therefore, participants not only spontaneously used the resource-abandonment option to communicate their benign intent, they also interpreted other players' use of this option as a signal of benign intent. Combining these experimental results with the evolutionary game analysis, we can conclude that *intSIG* is not only theoretically but also empirically viable as a strategy for indirect reciprocity. In future research, however, we need to verify whether the equilibrium of *intSIG* players spontaneously emerges when real participants play with each other. As for the standing strategy, participants did not use second-order information in experiment 1, but used it to some extent in experiment 2. Participants in experiment 2 might have learned the standing strategy from the series of feedback they received while in the role of recipient.

The *intSIG* strategy fosters cooperation by allowing players to signal their intention in a costly manner. This might appear to be largely deviating from the traditional approach to indirect reciprocity. However, its implication has much in common with some recent literature on indirect reciprocity. In Ghang & Nowak's model [38], each player can first decide whether to interact with the current partner. Declining interactions with uncooperative players does not hurt cooperators' reputation. In similar vein, Roberts [39] added the option of partner choice. Each donor can keep searching for a partner until he/she meets one whose image score satisfies his/her criterion. Although these strategies are different from *intSIG*, all seem to converge on a common theme. A key to stabilizing cooperation via indirect reciprocity is to give cooperative players some behavioural option that distinguishes their apparently uncooperative behaviours (e.g. not giving a resource to bad players) from genuinely uncooperative behaviours. Such an option may be any behaviour as far as there is no incentive for genuine defectors to perform it.

Despite the theoretical and empirical support for the viability of *intSIG* as a strategy for indirect reciprocity, it is not clear how the requisite signalling propensity and signal-reading ability co-evolved in the first place. One possibility is that they first co-evolved in a direct reciprocity context. When a pair of tit-for-tat players engage in the iterated prisoner's dilemma, even a single instance of careless defection leads to an endless alternating cycle of cooperation and defection [40]. Immediate communication of a careless defector's benign intent could therefore allow tit-for-tat players to avoid such futile alterations of cooperation/defection. Alternatively, these signalling mechanisms may have originated in a partner choice context. Unlike indirect reciprocity, where there is a cost associated with helping 'good' players, choosy players in a partner choice context do not have to incur cost of choosiness [21,41,42]. Accordingly, the signal-reading ability might have first evolved in the partner choice context. Moreover, when players can voluntarily initiate and terminate relationships, a costly signal of benign intent after an implementation error could prevent the premature dissolution of potentially beneficial, long-term relationships [32]. Admittedly, we have no decisive answer regarding under which context the signalling system first emerged. However, once evolved in some domain, it might have been exapted to the indirect reciprocity context.

A broader implication of this study is concerned with the importance of signalling behaviours in human cooperation. The theory of competitive altruism already linked signalling behaviours to cooperation [43,44]. However, the theory conceptualizes altruistic behaviours themselves as signals. On the other hand, it has been documented that many apparently wasteful behaviours, which cannot be equated with altruistic behaviours, also serve as commitment signals and facilitate dyadic cooperation by cementing interpersonal bonds [45–47]. The *intSIG* strategy likewise incorporates a signalling option independent of cooperation, and allows players to maintain their good standing even when they withhold help. This idea is resonant with the notion of communicative cooperation coined by Nöe [48]. Although it was proposed to underscore the importance of communication in animal cooperation, communications via signals should be no less important for human beings as we are not only a highly cooperative species but also an extremely communicative one. Therefore, supplementing traditional dichotomous behavioural options (cooperate and defect) with signals in evolutionary game models seems necessary to fully understand human sociality.

Ethics. This study was approved by the institutional review board at the corresponding author's institute.

Data accessibility. The data used in the reported analyses have been uploaded to the Dryad Digital Repository.

Authors' contributions. H.T. conducted the experiment, analysed the data, wrote the relevant part of the manuscript and approved the final version of the manuscript. H.O. conducted the game analyses, wrote the relevant part of the manuscript and approved the final version of the manuscript. Y.O. designed the experiment and wrote the final version of the manuscript.

Competing interests. We have no competing interests.

Funding. This study was generously supported by the Japan Society for the Promotion of Science KAKENHI grants to H.T. (15J05541), H.O. (25118006) and Y.O. (26590132, 15H03447), and by the John Templeton Foundation.

Acknowledgements. We are grateful to Naoki Konishi, Keisuke Matsugasaki, Adam Smith, Ayano Yagi, Chiaki Yamaguchi, Mana Yamaguchi and Ye-Yun Yu for their assistance.

References

1. Bowles S, Gintis H. 2011 *A cooperative species*. Princeton, NJ: Princeton University Press.
2. Fehr E, Fischbacher U. 2003 The nature of human altruism. *Nature* **425**, 785–791. (doi:10.1038/nature02043)
3. Nowak MA, Highfield R. 2012 *SuperCooperators*. New York, NY: Free Press.
4. Trivers R. 1971 The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35–57. (doi:10.1086/406755)
5. Axelrod R, Hamilton WD. 1981 The evolution of cooperation. *Science* **211**, 1390–1396. (doi:10.1126/science.7466396)
6. Alexander R. 1987 *The biology of moral systems*. New York, NY: Aldine de Gruyter.
7. Nowak MA, Sigmund K. 1998 Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573–577. (doi:10.1038/31225)
8. Nowak MA, Sigmund K. 1998 The dynamics of indirect reciprocity. *J. Theor. Biol.* **194**, 561–574. (doi:10.1006/jtbi.1998.0775)

9. Nowak MA, Sigmund K. 2005 Evolution of indirect reciprocity. *Nature* **437**, 1291–1298. (doi:10.1038/nature04131)
10. Leimar O, Hammerstein P. 2001 Evolution of cooperation through indirect reciprocity. *Proc. R. Soc. Lond. B* **268**, 745–753. (doi:10.1098/rspb.2000.1573)
11. Panchanathan K, Boyd R. 2003 A tale of two defectors: the importance of standing for evolution of indirect reciprocity. *J. Theor. Biol*. **224**, 115–126. (doi:10.1016/S0022-5193(03)00154-1)
12. Ohtsuki H, Iwasa Y. 2004 How should we define goodness? Reputation dynamics in indirect reciprocity. *J. Theor. Biol*. **231**, 107–120. (doi:10.1016/j.jtbi.2004.06.005)
13. Ohtsuki H, Iwasa Y. 2006 The leading eight: social norms that can maintain cooperation by indirect reciprocity. *J. Theor. Biol*. **239**, 435–444. (doi:10.1016/j.jtbi.2005.08.008)
14. Engelmann D, Fischbacher U. 2009 Indirect reciprocity and strategic reputation building in an experimental helping game. *Games Econ. Behav*. **67**, 399–407. (doi:10.1016/j.geb.2008.12.006)
15. Malle BF. 2004 *How the mind explains behavior*. Cambridge, MA: MIT Press.
16. Herrmann E, Call J, Hernández-Lloreda MV, Hare B, Tomasello M. 2007 Humans have evolved specialized skills of social cognition: the cultural intelligence hypothesis. *Science* **317**, 1360–1366. (doi:10.1126/science.1146282)
17. Ohtsubo Y, Rapoport A. 2006 Depth of reasoning in strategic form games. *J. Socio-Econ*. **35**, 31–47. (doi:10.1016/j.socec.2005.12.003)
18. Milinski M, Semmann D, Bakker TCM, Krambeck H-J. 2001 Cooperation through indirect reciprocity: Image scoring or standing strategy? *Proc. R. Soc. Lond. B* **268**, 2495–2501. (doi:10.1098/rspb.2001.1809)
19. Ule A, Schram A, Riedl A, Cason TN. 2009 Indirect punishment and generosity toward strangers. *Science* **326**, 1701–1704. (doi:10.1126/science.1178883)
20. Swakman V, Molleman L, Ule A, Egas M. 2016 Reputation-based cooperation: empirical evidence for behavioral strategies. *Evol. Hum. Behav*. **37**, 230–235. (doi:10.1016/j.evolhumbehav.2015.12.001)
21. Raihani NJ, Bshary R. 2015 Third-party punishers are rewarded, but third-party helpers even more so. *Evolution* **69**, 993–1003. (doi:10.1111/evo.12637)
22. Barclay P, Willer R. 2007 Partner choice creates competitive altruism in humans. *Proc. R. Soc. B* **274**, 749–753. (doi:10.1098/rspb.2006.0209)
23. Milinski M, Semmann D, Krambeck H-J. 2002 Reputation helps solve the 'tragedy of the commons'. *Nature* **415**, 424–426. (doi:10.1038/415424a)
24. Semmann D, Krambeck H-J, Milinski M. 2004 Strategic investment in reputation. *Behav. Ecol. Sociobiol*. **56**, 248–252. (doi:10.1007/s00265-004-0782-9)
25. Bereczkei T, Birkas B, Kerekes Z. 2007 Public charity offer as a proximate factor of evolved reputation-building strategy: an experimental analysis of a real-life situation. *Evol. Hum. Behav*. **28**, 277–284. (doi:10.1016/j.evolhumbehav.2007.04.002)
26. Bereczkei T, Birkas B, Kerekes Z. 2010 Altruism towards strangers in need: costly signaling in an industrial society. *Evol. Hum. Behav*. **31**, 95–103. (doi:10.1016/j.evolhumbehav.2009.07.004)
27. Yoeli E, Hoffman M, Rand DG, Nowak MA. 2013 Powering up with indirect reciprocity in a large-scale field experiment. *Proc. Natl Acad. Sci. USA* **110**, 10 424–10 429. (doi:10.1073/pnas.1301210110)
28. Carlson M, Miller N. 1987 Explanation of the relation between negative mood and helping. *Psychol. Bull*. **102**, 91–108. (doi:10.1037/0033-2909.102.1.91)
29. Salovey P, Mayer JD, Rosenhan DL. 1991 Mood and helping: mood as a motivator of helping and helping as a regulator of mood. In *Prosocial behavior* (ed. MS Clark), pp. 215–237. Newbury Park, CA: Sage.
30. Goffman E. 1971 *Relations in public: microstudies of the public order*. New York, NY: Basic Books.
31. Schlenker BR, Darby BW. 1981 The use of apologies in social predicaments. *Soc. Psychol. Q*. **44**, 271–278. (doi:10.2307/3033840)
32. Ohtsubo Y, Watanabe E. 2009 Do sincere apologies need to be costly? Test of a costly signaling model of apology. *Evol. Hum. Behav*. **30**, 114–123. (doi:10.1016/j.evolhumbehav.2008.09.004)
33. Watanabe E, Ohtsubo Y. 2012 Costly apology and self-punishment after an unintentional transgression. *J. Evol. Psychol*. **10**, 87–105. (doi:10.1556/JEP.10.2012.3.1)
34. Inbar Y, Pizarro DA, Gilovich T, Ariely D. 2013 Moral masochism: on the connection between guilt and self-punishment. *Emotion* **13**, 14–18. (doi:10.1037/a0029749)
35. Nelissen RMA, Zeelenberg M. 2009 When guilt evokes self-punishment: evidence for the existence of a Dobby effect. *Emotion* **9**, 118–122. (doi:10.1037/a0014540)
36. Tanaka H, Yagi A, Komiya A, Mifune N, Ohtsubo Y. 2015 Shame-prone people are more likely to punish themselves: a test of the reputation-maintenance explanation for self-punishment. *Evol. Behav. Sci*. **9**, 1–7. (doi:10.1037/ebs0000016)
37. Wedekind C, Milinski M. 2000 Cooperation through image scoring in humans. *Science* **288**, 850–852. (doi:10.1126/science.288.5467.850)
38. Ghang W, Nowak MA. 2015 Indirect reciprocity with optional interactions. *J. Theor. Biol*. **365**, 1–11. (doi:10.1016/j.jtbi.2014.09.036)
39. Roberts G. 2015 Partner choice drives the evolution of cooperation via indirect reciprocity. *PLoS ONE* **10**, e0129442. (doi:10.1371/journal.pone.0129442)
40. Nowak MA, Sigmund, K. 1992 Tit for tat in heterogeneous populations. *Nature* **355**, 250–253. (doi:10.1038/355250a0)
41. Barclay P. 2013 Strategies for cooperation in biological markets, especially for humans. *Evol. Hum. Behav*. **34**, 164–175. (doi:10.1016/j.evolhumbehav.2013.02.002)
42. Sylwester K, Roberts G. 2013 Reputation-based partner choice is an efficient alternative to indirect reciprocity in solving social dilemmas. *Evol. Hum. Behav*. **34**, 201–206. (doi:10.1016/j.evolhumbehav.2012.11.009)
43. Zahavi A, Zahavi A. 1997 *The handicap principle: a missing piece of Darwin's puzzle*. New York, NY: Oxford University Press.
44. Roberts G. 1998 Competitive altruism: from reciprocity to the handicap principle. *Proc. R. Soc. Lond. B* **265**, 427–431. (doi:10.1098/rspb.1998.0312)
45. Frank RH. 1988 *Passions within reason: the strategic role of the emotions*. New York, NY: Norton.
46. Nesse RM (ed). 2001 *Evolution and the capacity for commitment*. New York, NY: Russell Sage Foundation.
47. Yamaguchi M, Smith A, Ohtsubo Y. 2015 Commitment signals in friendship and romantic relationships. *Evol. Hum. Behav*. **36**, 467–474. (doi:10.1016/j.evolhumbehav.2015.05.002)
48. Noë R. 2006 Cooperation experiments: coordination through communication versus acting apart together. *Anim. Behav*. **71**, 1–18. (doi:10.1016/j.anbehav.2005.03.037)