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

Tanaka, Hiroki
Ohtsuki, Hisashi
Ohtsubo, Yohsuke

(Citation)
Proceedings of the Royal Society B: Biological Sciences,283(1835):20160694-20160694

(Issue Date)
2016-07-27

(Resource Type)
journal article

(Version)
Accepted Manuscript

The Price of Being Seen to Be Just:

An Intention Signalling Strategy for Indirect Reciprocity

Hiroki Tanaka[1, 2] Hisashi Ohtsuki[3] 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

[3]School of Advanced Sciences, SOKENDAI (The Graduate University for Advanced Studies), Shonan Village, Hayama 240-0193, Kanagawa, Japan

**author for correspondence (yohtsubo@lit.kobe-u.ac.jp).*

Abstract

Cooperation amongst 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', but not '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 utilise this option, they also interpreted others' use of this option as a signal of cooperative intent.

Keywords: indirect reciprocity; cooperation; intention signalling; mind-reading

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 [6, 7]. 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 payoff can be maintained by only eight (out of 4,096) strategies, and that all of them, collectively called the 'leading eight', distinguish justified from unjustified defection [12, 13]. It is important to notice 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 stabilising 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 stabilise cooperation by indirect reciprocity. We then reports 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. Notice 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 utilise second-order information. Instead, it assumes the co-evolution 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 materials 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 payoff 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 payoff is $b/(2-\omega)$. Therefore, the comparison between $(b-c)/[2(1-\omega)]$ and $b/(2-\omega)$ reveals that *intSIG*'s net payoff exceeds *ALLD*'s net payoff if ω is sufficiently large: $\omega > 2c/(b+c)$.

In contrast, *ALLC* players obtain the same net payoff 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 payoff difference arises between *intSIG* and *ALLC*, and *intSIG* can resist invasion by *ALLC*. Let us pay attention to this payoff difference. After committing an implementation error as a donor, *intSIG* abandons c to produce the signal. Therefore, the payoff 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 payoff consequences. In 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 stabilises *intSIG* against *ALLC* is identical to the condition that stabilises *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 operationalisation 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), '*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 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 ≈ £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 programme with a small probability. If errors occurred, participants in the donor role were made immediately aware. Unbeknownst to participants, 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 materials).

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 (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 materials 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 1a, 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 $p < .017$). 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 (.341, $sd = .101$) than unjustified defectors (.007, $sd = .000$): $t_{30} = 3.77$, $p < .001$.

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 2a. As expected, the main effect of partner type was significant ($F_{2, 102} = 34.83$, $p < .001$), 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 ($F_{3, 153} = 20.49$, $p < .001$), a post-hoc test by Ryan's method indicated that participants gave the resource to GG- and GN-recipients more

frequently than NG- and NN-recipients (figure 3a).

(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 1b). 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 ($p < .001$ for each comparison). On the other hand, signalling frequency after implementation error and justified defection did not differ ($p = .293$). 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 5 participants were identified as unjustified defectors. These 5 unjustified defectors (.48, $sd = .17$) used the signal option less frequently than the 18 justified defectors (.82, $sd = .08$), $t_{21} = 2.14$, $p = .044$. 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 2b). The effect of recipient type was significant ($F_{2, 96} = 31.38$, $p < .001$), 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-recipient 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 utilised 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 as 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 and 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 the 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 stabilising 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 prisoners' 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 conceptualises 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

333 important for human beings as we are not only a highly cooperative species but also an extremely
334 communicative one. Therefore, supplementing traditional dichotomous behavioural options
335 (cooperate and defect) with signals in evolutionary game models seems necessary to fully understand
336 human sociality.
337
338
339 **Ethics statement.** This study was approved by the institutional review board at the corresponding
340 author's institute.
341 **Data accessibility statement.** The data used in the reported analyses have been uploaded to the
342 Dryad Digital Repository.
343 **Competing interests statement.** We have no competing interests.
344 **Authors' contributions statement.** H. Tanaka conducted the experiment, analysed the data, wrote
345 the relevant part of the manuscript and approved the final version of the manuscript. H. Ohtsuki
346 conducted the game analyses, wrote the relevant part of the manuscript, and approved the final
347 version of the manuscript. Y. Ohtsubo designed the experiment, wrote the final version of the
348 manuscript.
349 **Acknowledgements.** We are grateful to Naoki Konishi, Keisuke Matsugasaki, Adam Smith, Ayano
350 Yagi, Chiaki Yamaguchi, Mana Yamaguchi, and Ye-Yun Yu for their assistance.
351 **Funding statement.** This study was generously supported by the Japan Society for the Promotion of
352 Science KAKENHI Grants to HT (15J05541), HO (25118006), and YO (26590132, 15H03447), and
353 by the John Templeton Foundation.

354 **References**

355 1. Bowles S, Gintis, H. 2011 *A cooperative species*. Princeton, NJ: Princeton University Press.

356 2. Fehr E, Fischbacher U. 2003 The nature of human altruism. *Nature* **425**, 785-791.
357 (doi:10.1038/nature02043)

358 3. Nowak MA, Highfield R. 2012 *SuperCooperators*. New York: Free Press.

359 4. Trivers R. 1971 The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35-57.
360 (doi:10.1086/406755)

361 5. Axelrod R, Hamilton WD. 1981 The evolution of cooperation. *Science* **211**, 1390-1396.
362 (doi:10.1126/science.7466396)

363 6. Alexander R. 1987 *The biology of moral systems*. New York: Aldine de Gruyter.

364 7. Nowak MA, Sigmund K. 1998 Evolution of indirect reciprocity by image scoring. *Nature* **393**,
365 573-577. (doi:10.1038/31225)

366 8. Nowak MA, Sigmund K. 1998 The dynamics of indirect reciprocity. *J. Theor. Biol.* **194**, 561-574.
367 (doi:10.1006/jtbi.1998.0775)

368 9. Nowak MA, Sigmund K. 2005 Evolution of indirect reciprocity. *Nature* **437**, 1291-1298.
369 (doi:10.1038/nature04131)

370 10. Leimar O, Hammerstein P. 2001 Evolution of cooperation through indirect reciprocity. *Proc. R.*
371 *Soc. Lond. B* **268**, 745-753. (doi:10.1098/rspb.2000.1573)

372 11. Panchanathan K, Boyd R. 2003 A tale of two defectors: The importance of standing for evolution
373 of indirect reciprocity. *J. Theor. Biol.* **224**, 115-126. (doi:10.1016/S0022-5193(03)00154-1)

374 12. Ohtsuki H, Iwasa Y. 2004 How should we define goodness?—Reputation dynamics in indirect
375 reciprocity. *J. Theor. Biol.* **231**, 107-120. (doi:10.1016/j.jtbi.2004.06.005)

376 13. Ohtsuki H, Iwasa Y. 2006 The leading eight: Social norms that can maintain cooperation by
377 indirect reciprocity. *J. Theor. Biol.* **239**, 435-444. (doi:10.1016/j.jtbi.2005.08.008)

378 14. Engelmann D, Fischbacher U. 2009 Indirect reciprocity and strategic reputation building in an

379 experimental helping game. *Games Econ. Behav.* **67**, 399-407. (doi:10.1016/j.geb.2008.12.006)
380 15. Malle BF. 2004 *How the mind explains behavior*. Cambridge, MA: MIT Press.
381 16. Herrmann E, Call J, Hernández-Lloreda MV, Hare B, Tomasello M. 2007 Humans have evolved
382 specialized skills of social cognition: The cultural intelligence hypothesis. *Science* **317**, 1360-
383 1366. (doi:10.1126/science.1146282)
384 17. Ohtsubo Y. Rapoport A. 2006 Depth of reasoning in strategic form games. *J. Socio. Econ.* **35**, 31-
385 47. (doi:10.1016/j.socec.2005.12.003)
386 18. Milinski M, Semmann D, Bakker TCM, Krambeck H-J. 2001 Cooperation through indirect
387 reciprocity: Image scoring or standing strategy? *Proc. R. Soc. Lond. B* **268**, 2495-2501.
388 (doi:10.1098/rspb.2001.1809)
389 19. Ule A, Schram A, Riedl A, Cason TN. 2009 Indirect punishment and generosity toward strangers.
390 *Science* **326**, 1701-1704. (doi:10.1126/science.1178883)
391 20. Swakman V, Molleman L, Ule A, Egas M. 2016 Reputation-based cooperation: Empirical
392 evidence for behavioral strategies. *Evol. Hum. Behav.* **37**, 230-235.
393 (doi:10.1016/j.evolhumbehav.2015.12.001)
394 21. Raihani, NJ, Bshary R. 2015 Third-party punishers are rewarded, but third-party helpers even
395 more so. *Evolution* **69**, 993-1003. (doi:10.1111/evo.12637)
396 22. Barclay P, Willer R. 2007 Partner choice creates competitive altruism in humans. *Proc. R. Soc. B*
397 **274**, 749-753. (doi:10.1098/rspb.2006.0209)
398 23. Milinski M, Semmann D, Krambeck H-J. 2002 Reputation helps solve the ‘tragedy of the
399 commons.’ *Nature* **415**, 424-426. (doi:10.1038/415424a)
400 24. Semmann D, Krambeck H-J, Milinski M. 2004 Strategic investment in reputation. *Behav. Ecol.*
401 *Sociobiol*. **56**, 248-252. (doi:10.1007/s00265-004-0782-9)
402 25. Bereczkei T, Birkas B, Kerekes Z. 2007 Public charity offer as a proximate factor of evolved
403 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. U.S.A.* **110**, 10424-10429. (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: Mirostudies of the public order*. New York: 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)

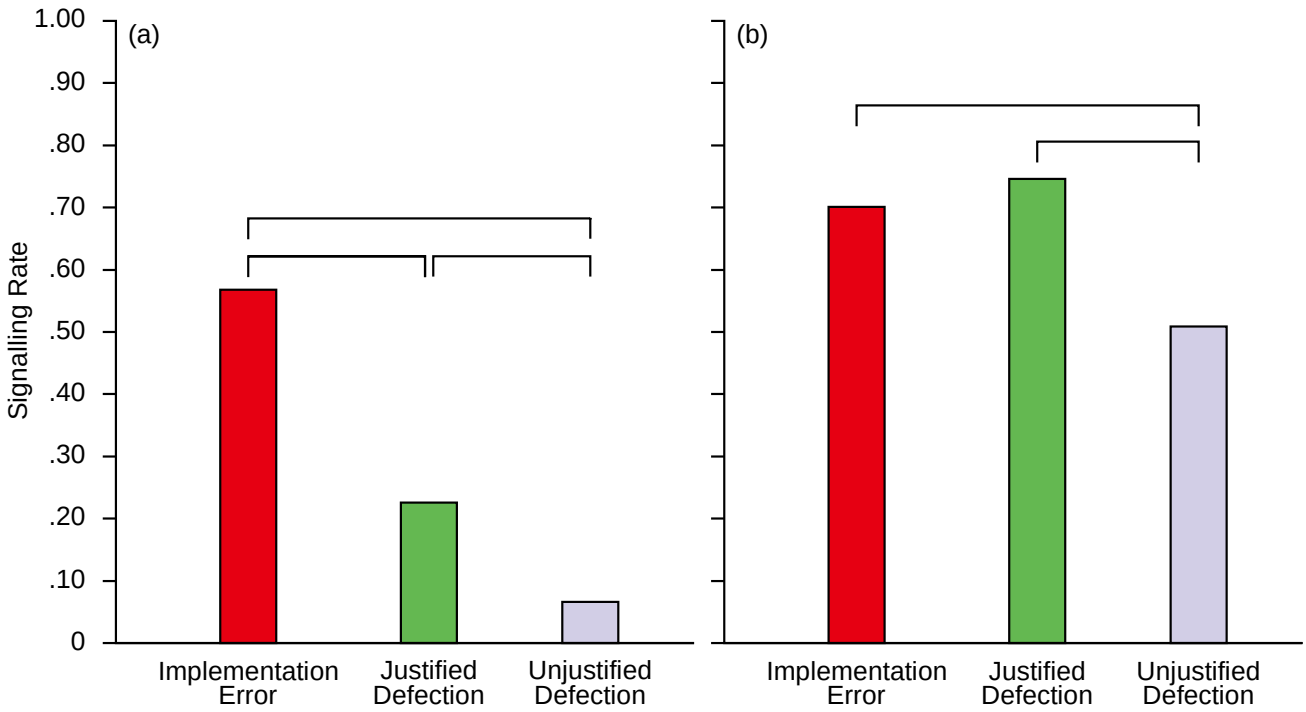
429 37. Wedekind C, Milinski M. 2000 Cooperation through image scoring in humans. *Science* **288**, 850-
430 852. (doi:10.1126/science.288.5467.850)
431 38. Ghang W, Nowak MA. 2015 Indirect reciprocity with optional interactions. *J. Theor. Biol.* **365**, 1-
432 11. (doi:10.1016/j.jtbi.2014.09.036)
433 39. Roberts G. 2015 Partner choice drives the evolution of cooperation via indirect reciprocity. *PLoS*
434 *ONE* **10**(6): e0129442. (doi:10.1371/journal.pone.0129442)
435 40. Nowak MA, Sigmund, K. 1992 Tit for tat in heterogeneous populations. *Nature* **355**, 250-253.
436 (doi:10.1038/355250a0)
437 41. Barclay P. 2013 Strategies for cooperation in biological markets, especially for humans. *Evol.*
438 *Hum. Behav.* **34**, 164-175. (doi:10.1016/j.evolhumbehav.2013.02.002)
439 42. Sylwester K, Roberts G. 2013 Reputation-based partner choice is an efficient alternative to
440 indirect reciprocity in solving social dilemmas. *Evol. Hum. Behav.* **34**, 201-206.
441 (doi:10.1016/j.evolhumbehav.2012.11.009)
442 43. Zahavi A, Zahavi A. 1997 *The handicap principle: A missing piece of Darwin's puzzle*. New
443 York: Oxford University Press.
444 44. Roberts G. 1998 Competitive altruism: From reciprocity to the handicap principle. *Proc. R. Soc.*
445 *Lond. B* **265**, 427-431 (doi:10.1098/rspb.1998.0312)
446 45. Frank RH. 1988 *Passions within reason: The strategic role of the emotions*. New York: Norton.
447 46. Nesse RM (ed). 2001 *Evolution and the capacity for commitment*. New York: Russell Sage
448 Foundation.
449 47. Yamaguchi M, Smith A, Ohtsubo Y. 2015 Commitment signals in friendship and romantic
450 relationships. *Evol. Hum. Behav.* **36**, 497-474. (doi:10.1016/j.evolhumbehav.2015.05.002)
451 48. Noë R. 2006 Cooperation experiments: Coordination through communication versus acting apart
452 together. *Anim. Behav.* **71**, 1-18. (doi:10.1016/j.anbehav.2005.03.037)

Figure Captions

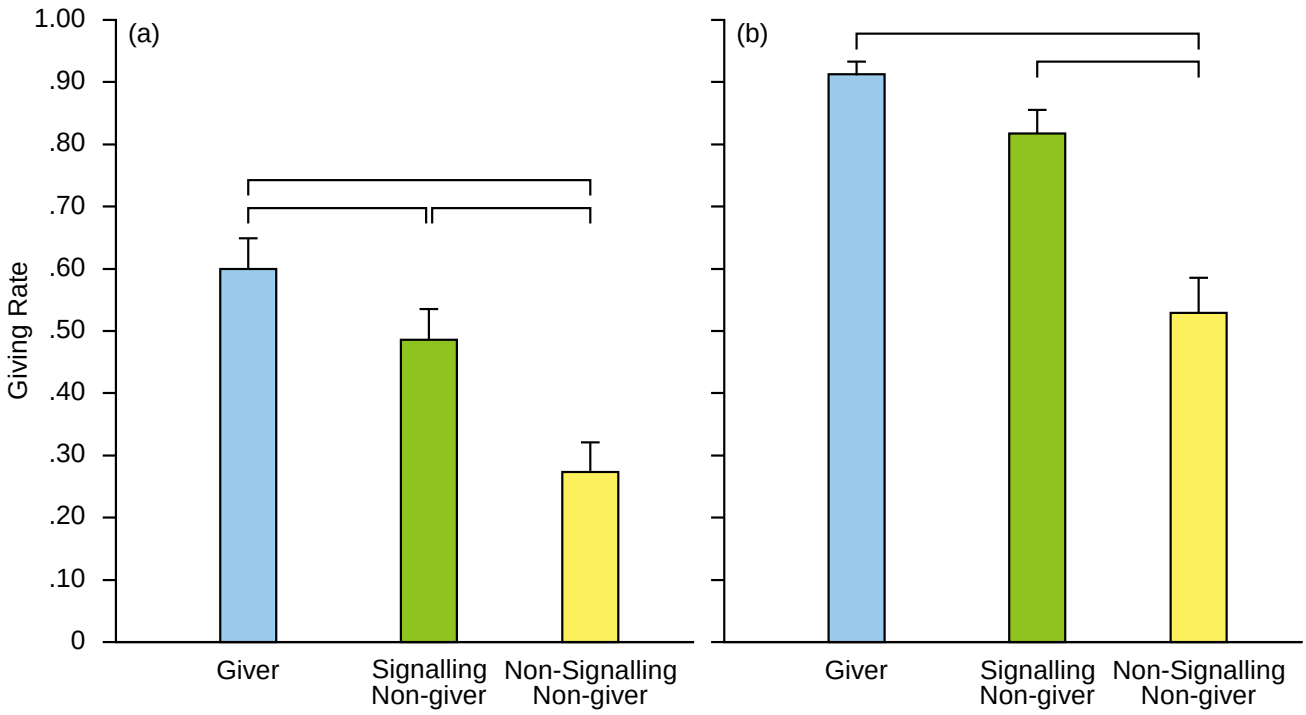
figure 1. Relative frequency of signal use after implementation error, justified defection, and unjustified defection in experiment 1 (a) and experiment 2 (b).

figure 2. Mean giving rate to giver, signalling non-giver, and non-signalling non-giver in the signalling condition of experiment 1 (a) and experiment 2 (b).

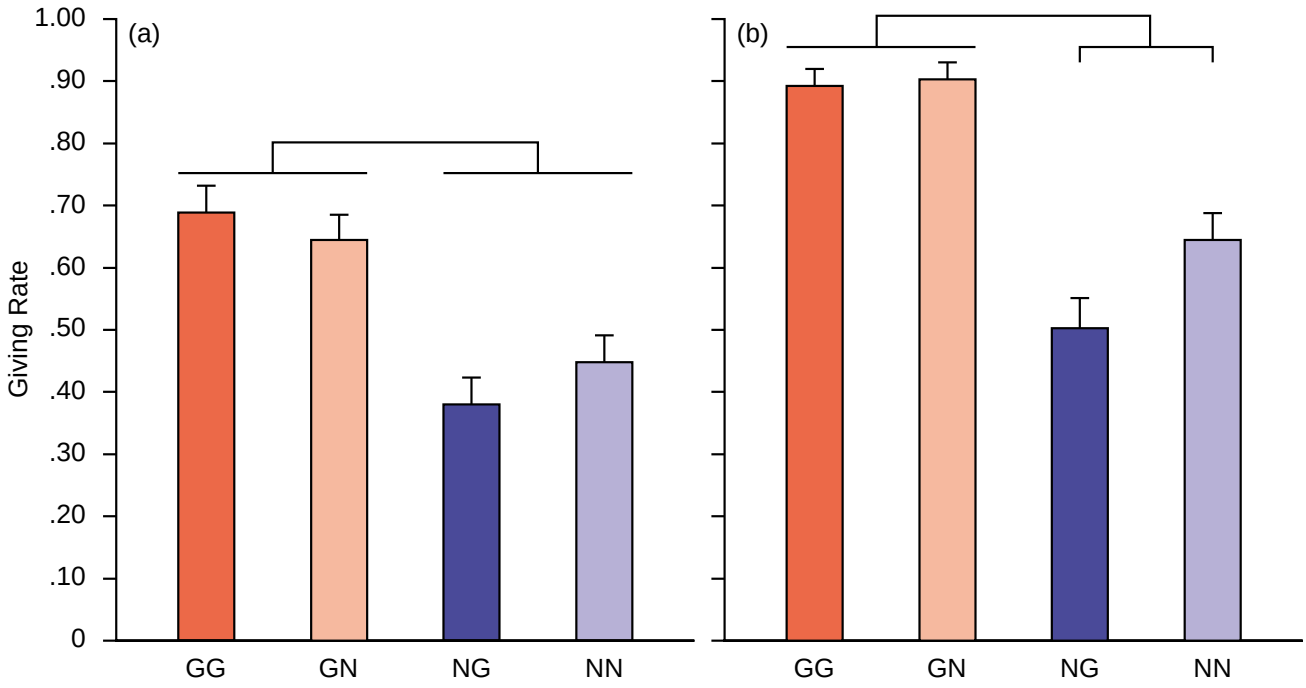
figure 3. Mean giving rate to GG, GN, NG, and NN recipient in the standing condition of experiment 1 (a) and experiment 2 (b).


(a)
(b)
Signalling Rate
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
0
Implementation Error
Justified Defection
Unjustified Defection
Implementation Error
Justified Defection
Unjustified Defection

<Figure 1>


(a)
(b)
Giving Rate
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
0
Giver
Signalling Non-giver
Non-Signalling Non-giver
Giver
Signalling Non-giver
Non-Signalling Non-giver

<Figure 2>


(a)
(b)
Giving Rate
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
0
GG
GN
NG
NN
GG
GN
NG
NN

<Figure 3>

Electronic Supplementary Materials

The Price of Being Seen to Be Just:

An Intention Signalling Strategy for Indirect Reciprocity

Hiroki Tanaka Hisashi Ohtsuki Yohsuke Ohtsubo

This document includes:

Evolutionary Game Analyses

(a) Payoff of *intSIG* strategy

Consider that every player from an infinitely large population is paired with another player and assigned either the donor or recipient role with the same probability in each round. A donor decides whether to help his recipient incurring a cost of c (cooperation) or not (defection). There is a small chance of implementation error (e) whereby each donor fails to help his recipient against his will. There is no possibility of erroneous cooperation (a donor who intends to defect but unintentionally cooperates). A recipient receives a benefit of b if her donor helps her, but receives nothing if her donor decides not to help or commits an implementation error. In addition to this standard indirect reciprocity game, the donor has another behavioural option, signalling, that is available only after intentional defection or implementation error. The donor is allowed to pay a cost of s to produce a signal. The *intSIG* strategy assigns a 'bad' standing only to a partner whose previous behaviour was 'defection without signal'. In other words, at the beginning of the game, each player is in good standing, and remains in good standing unless his previous choice was 'defection without signal'. The payoffs of the two players in each round and the donor's post-round standing are summarised as Table S1. After playing t-th round, there is the $(t+1)$th round with the probability of ω for all $t=1, 2,\ldots$ In each round, each player will be randomly matched with a new partner, and is assigned either the donor or recipient role with the probability of 0.5.

Table S1

The payoff of donor and recipient as a function of donor behaviour

Donor's Behaviour	Donor	Recipient	**Donor's Post-round Standing**
Cooperation	$-c$	b	**good**
Defection with Signal	$-s$	0	**good**
Defection without Signal	0	0	**bad**

We first computed *intSIG*'s payoff. When the entire group consists of *intSIG* players, in each round, an *intSIG* player earns $(1-e)(-c)+e(-s)$ as a donor (he cooperates with the probability of $1-e$, while failing to do so and producing the signal with the probability of e). Regardless of implementation error, *intSIG* is always in good standing because it uses the signal option. Therefore, *intSIG* earns $(1-e)(b)$ as a recipient in each round. Because the donor and recipient roles are assigned with the probability of 0.5, *intSIG* earns on average w_{SIG} in each round:

$$w_{SIG} = \frac{(1-e)(b-c)-es}{2}. \quad (1)$$

As this game continues with the probability of ω, the net payoff of *intSIG*, W_{SIG}, is written as follows:

$$W_{SIG} = \frac{1}{1-\omega}\frac{(1-e)(b-c)-es}{2}. \quad (2)$$

It is easy to compare the net payoff of the *intSIG* group with that of the *ALLD* group and the *ALLC* group. When all group members are unconditional defectors (*ALLD*), donors never help their recipients, and recipients never receive any benefit. Accordingly, $w_{ALLD} = 0$. Therefore, at the entire group level, the *intSIG* group is more profitable than *ALLD* when the following condition holds:

$$\frac{(1-e)(b-c)-es}{2} > 0,$$

which is reduced as $e < \frac{b-c}{b-c+s}$.

This condition indicates that *intSIG* is more profitable than *ALLD* unless the error rate is large and/or the signal cost is high. For example, if $b = 1.5$, $c = s = 1$, the error rate only needs to be smaller than .33.

If a group consists of only unconditional cooperators (*ALLC*), each player incurs a cost of c as a donor unless he commits the implementation error with the probability of e. An *ALLC* player will receive the benefit of b as a recipient unless her partner commits the error with the probability of e. Accordingly, *ALLC* earns $b-c$ with the probability of $1-e$, and earns 0 with the probability of e. Thus, each *ALLC* player earns

$$w_{ALLC} = \frac{(1-e)(b-c)}{2}. \quad (3)$$

If we compare the payoffs of *ALLC* and *intSIG*, it is evident that *intSIG* cannot outperform *ALLC* as a group because the following condition (4) is inconsistent with our assumption that $e > 0$ and $s > 0$.

$$\frac{(1-e)(b-c)-es}{2} > \frac{(1-e)(b-c)}{2}. \quad (4)$$

In sum, a group of *intSIG* players can outperform a group of *ALLD* players unless they are highly likely to commit errors. On the other hand, a group of *intSIG* players can never outperform a group of *ALLC* players. This is because *ALLC* will never waste their resource by producing a signal. However, because of their unconditional cooperativeness, they are easily exploited by an uncooperative strategy. In the next section, we examine whether *intSIG* is stable against the invasion of an exploitative strategy, *ALLD*, and whether *intSIG* is vulnerable to the invasion of a cooperative strategy, *ALLC*.

(b) Evolutionary Stability of *intSIG* against *ALLD*

We then examined the condition under which rare *ALLD* players cannot invade a group of *intSIG* players. We first computed the expected payoff of a rare *ALLD* in a group of *intSIG* players. Because we assume the frequency of *ALLD* is negligible, the net payoff of *intSIG* players is written as Eq. (2).

When *ALLD* is a donor, it earns 0 as it does not help the partner. When *ALLD* is a recipient, it earns either $(1-e)b$ if its standing is good or 0 if its standing is bad. Let $G_{ALLD}(t)$ be the probability that *ALLD* is in good standing after *t*-th round. Under the assumption of the model that each player starts with a good standing, $G_{ALLD}(0) = 1$, *ALLD*'s standing becomes 'bad' once it plays the role of donor, and never returns to 'good'. Because the donor role is assigned with the probability of 0.5, an *ALLD* player in good standing will shift to bad standing with the probability of 0.5. Therefore, $G_{ALLD}(t+1) = G_{ALLD}(t)\times(1/2)$.

Accordingly,

$$G_{ALLD}(t) = \left(\frac{1}{2}\right)^t . \quad (5)$$

ALLD's payoff in the *t*-th round, $w_{ALLD}(t)$, is the product of the probability of being in good standing, the probability of being assigned to the recipient role, and the benefit conferred by a cooperative opponent (the payoff in the donor role is always 0, and can be ignored):

$$w_{ALLD}(t) = \left(\frac{1}{2}\right)^{t-1} \frac{1}{2}(1-e)b = \left(\frac{1}{2}\right)^t (1-e)b . \quad (6)$$

Because the game continues with the probability of ω, the net payoff of *ALLD* is:

$$W_{ALLD} = \sum_{t=1}^{\infty} w_{ALLD}(t) = \frac{1}{2-\omega}(1-e)b. \quad (7)$$

Based on Eq. (2) and Eq. (7), *ALLD* cannot invade a group of *intSIG* players as far as the following condition holds:

$$W_{SIG} > W_{ALLD}$$

$$\Leftrightarrow \frac{1}{1-\omega}\frac{(1-e)(b-c)-es}{2} > \frac{1}{2-\omega}(1-e)b$$

$$\Leftrightarrow \frac{1-e}{e}\{(2-\omega)(b-c) - 2(1-\omega)b\} > s(2-\omega) \quad . \quad (8)$$

In the above condition (8), when the assumption that the error rate (e) is small, $\frac{1-e}{e}$ takes a large positive value. Also, the right side of the inequality is always positive (both s and $2-\omega$ take positive values). Therefore, it is expected that Inequality (8) holds if $(2-\omega)(b-c) - 2(1-\omega)b > 0$. This condition is rewritten as follows:

$$\omega > \frac{2c}{b+c} \quad . \quad (9)$$

Based on the assumption that $b > c$, the range of the right side of Inequality (9) is $0 < \frac{2c}{b+c} < 1$, which corresponds with the range of ω. Accordingly, Inequality (9) reveals that, when the implementation error rate e is tiny but positive, *intSIG* is stable against *ALLD* as far as the game continues with the probability larger than $\frac{2c}{b+c}$. For example, when $b = 2$ and $c = 1$, condition (9) only requires that the games consist of more than 3 rounds on average ($\omega > 2/3$). It is important to notice that this condition does not depend on the size of signal cost, s.

In the main text, we assume that the signal cost, *s*, is equal to the cost of cooperation, *c*. Substituting *s* in Inequality (8) with *c* yields the following condition:

$$e < 1 - \frac{(2-\omega)c}{\omega b} \quad . \tag{10}$$

This condition holds when the game continues for a substantially long period of time. For example, when ω is nearly 1, this condition becomes $e < 1 - \frac{c}{b}$. Therefore, if the game continues for a substantially long time and *e* is sufficiently small, *ALLD* cannot invade the group of *intSIG* players.

(c) Evolutionary Stability of *intSIG* against *ALLC*

We explored the condition under which *intSIG* is stable against the invasion of *ALLC*. When there is no possibility of implementation errors, rare *ALLC* players and *intSIG* players will peacefully co-exist at the cooperative equilibrium. However, if the possibility of implementation errors is introduced, the payoffs of *intSIG* and *ALLC* will diverge because *intSIG* players can maintain their good standing by producing a costly signal, while *ALLC* players have to wait for one donor-round so that they can cooperate and restore their good standing.

To obtain the net payoff of *ALLC* in the *intSIG* group, let $G_{ALLC}(t)$ be the probability that *ALLC* is in good standing after *t*-th round. We have $G_{ALLC}(0) = 1$ as an initial condition. *ALLC*'s standing becomes 'bad' only when it commits an implementation error. Therefore, after playing the donor role, its standing is 'good' with the probability of $1-e$. After playing the recipient role, its standing does not change. Accordingly, the probability that *ALLC* is in good standing after $(t+1)$-th round is

$$G_{ALLC}(t+1) = \frac{1}{2} G_{ALLC}(t) + \frac{1-e}{2}. \tag{11}$$

Subtracting $1-e$ from both sides of Eq. (11) yields

$$G_{ALLC}(t+1) - (1-e) = \frac{1}{2} G_{ALLC}(t) - \frac{1-e}{2}. \tag{12}$$

Let $H_{ALLC}(t) = G_{ALLC}(t) - (1-e)$, and Eq. (12) can be rewritten as

$$H_{ALLC}(t+1) = \frac{1}{2} H_{ALLC}(t). \tag{13}$$

Notice that $H_{ALLC}(0) = 1-(1-e) = e$. Therefore,

$$H_{ALLC}(t) = G_{ALLC}(t) - (1-e) = e\left(\frac{1}{2}\right)^t. \quad (14)$$

From Eq. (14), we obtained the probability that *ALLC* is in good standing after *t*-th round as follows:

$$G_{ALLC}(t) = e\left(\frac{1}{2}\right)^t + (1-e). \quad (15)$$

Using Eq. (15), we can compute the expected payoff of *ALLC* at the *t*-th round. If *ALLC* plays the donor role, its payoff is $-c(1-e)$ regardless of its standing. If *ALLC* plays the recipient role, its expected payoff is $b(1-e)$ when its standing is good, while the expected payoff is 0 if its standing is bad. Accordingly, *ALLC*'s expected payoff at the *t*-th round is written as

$$w_{ALLC}(t) = -\frac{1}{2}(1-e)c + \frac{1}{2}b(1-e)G_{ALLC}(t-1)$$

$$= \left(\frac{1}{2}\right)^t e(1-e)b + \frac{1}{2}\{(1-e)^2 b - (1-e)c\}. \quad (16)$$

From Eq. (16), *ALLC*'s net payoff is derived as follows:

$$W_{ALLC} = \frac{1}{2-\omega}e(1-e)b + \frac{1}{1-\omega}\frac{(1-e)^2 b-(1-e)c}{2}. \quad (17)$$

Based on Eq. (2) and Eq. (17), the condition under which *intSIG* is stable against *ALLC* ($W_{SIG} > W_{ALLC}$) is derived as follows:

$$\frac{1}{1-\omega}\frac{(1-e)(b-c)-es}{2} > \frac{1}{2-\omega}e(1-e)b + \frac{1}{1-\omega}\frac{(1-e)^2 b-(1-e)c}{2},$$

which is rewritten as

$$(2-\omega)e(1-e)b - (2-\omega)es > 2(1-\omega)e(1-e)b. \quad (18)$$

By dividing the both sides of Inequality (18) by $e>0$, the ESS condition of *intSIG* against *ALLC* was further rewritten as below:

$$e < 1 - \frac{(2-\omega)s}{\omega b} \quad . \quad (19)$$

Because we divided both sides of inequality by a small number, *e*, to obtain the condition (19), the difference between the net payoffs of *intSIG* and *ALLC* is small. However, if condition (19) holds, *intSIG* is stable against *ALLC*. This tends to hold when the cost of the signal, *s*, is relatively small

compared to the benefit of being helped, *b*. In other words, unlike the ESS condition against *ALLD*, which did not depend on the cost of the signal, *intSIG* is less likely to be stable against *ALLC* if the signal cost is large.

We further examined condition (19) assuming that the signal cost, *s*, is equal to the cost of cooperation, *c*. Interestingly, the resultant condition was exactly equal to the condition under which *intSIG* was stable against the invasion of *ALLD*, which is condition (10)

$$e < 1 - \frac{(2-\omega)c}{\omega b} \quad . \tag{20}$$

(d) Summary

We investigated under what conditions *intSIG* is evolutionarily stable against *ALLD* and *ALLC*. First, *intSIG* was stable against *ALLD* as far as the interactions continue for a sufficiently long period of time, and the stability condition did not depend on the cost of the signal. Second, although *intSIG*'s and *ALLC*'s expected payoffs were close to each other, *intSIG* was stable against *ALLC* when the cost of the signal was not too large. When we assumed that the cost of the signal, *s*, was equal to the cost of cooperation, *c*, which is a sufficient amount of signalling cost to prevent dishonest signallers from undermining the separating equilibrium, it was shown that *intSIG* was stable against both *ALLD* and *ALLC* under exactly the same condition. Therefore, we can conclude that *intSIG* is robust against *ALLC*, which typically allows *ALLD*'s invasion.

II. Additional analyses of game behaviours

(a) Cooperation rate in the practice session

In the practice session, participants played the standard giving game. In both conditions, participants played the identical game, which gave neither second-order information nor the signalling option to participants. For each participant, we computed the mean cooperation rate towards the 'good' recipient (the recipient who chose 'give' in the previous round) and 'bad' recipient separately. A 2 (recipient type: good vs. bad) × 2 (game type: signalling vs. standing) ANOVA including the former factor as repeated measures indicated that only the main effect of recipient type was significant, $F_{1, 102} = 78.07$, $p < .001$, and other effects were not significant in experiment 1. Participants were more likely to give their resource to the 'good 'recipient (.71, *sd* = 0.30) than the 'bad' recipient (.45, *sd* = 0.26).

In experiment 2, where participants played the game against four image-scoring players and one *ALLD* player, the comparable ANOVA again revealed the significant main effect of recipient type, $F_{1, 97} = 58.73$, $p < .001$ (.83, *sd* = 0.22 vs. .65, *sd* = 0.26 towards the 'good' vs. 'bad' recipient, respectively). However, in experiment 2, an unexpected interaction effect between the recipient type and game type was also significant, $F_{1, 97} = 5.74$, $p = .019$. Participants were less likely to give the resource to the 'bad' recipient in the signalling condition (.58, *sd* = 0.29) than in the standing condition (.71, *sd* = 0.23). We do not have any good explanation for this unexpected effect as we randomly assigned participants to the two game conditions, and we did not give any condition-specific instructions at this stage (prior to the main signalling vs. standing game).

In sum, the results of the practice session clearly showed that participants discriminated recipients in terms of the recipients' previous behaviour. We thus proceeded to examine how participants' behaviour towards the previous giver and non-giver would be moderated by the opportunity of signalling or the availability of second-order information.

(b) Reaction time in the experimental games

According to Milinski et al. [1], the standing strategy, which utilises second-order information, is cognitively demanding and thus difficult for people to use. If *intSIG* is a more intuitive strategy than the standing strategy, it is predicted that the time to make the decision ('give' or 'not give') will be shorter in the signalling condition than in the standing condition. Therefore, we compared the reaction time (RT) in the two conditions. The prediction was corroborated only in experiment 2. In experiment 1, although the mean RT was slightly shorter in the signalling condition (2.61 sec., *sd* = 1.17) than in the standing condition (2.74 sec., *sd* = 0.87), the difference was not statistically significant, $t_{102} = 0.61$, $p = .54$. On the other hand, in experiment 2, the mean RT was significantly shorter in the signalling condition (2.19 sec., *sd* = 0.72) than in the standing condition (2.49 sec., *sd* = 0.76), $t_{102} = 1.99$, $p = .049$. Recall that participants in the standing condition did not utilise second-order information in experiment 1, whereas participants utilised second-order information in experiment 2. Therefore, the different pattern in the RT data might be explained by whether participants utilised second-order information. Although the results are not conclusive, information about the partner's behaviour plus signal appears less cognitively taxing than information about the partner's behaviours plus the partner's previous partner's behaviour.

(c) Total payoff in the experimental games

We then examined in which condition (signalling vs. standing) participants earned a greater net payoff. In experiment 1, the mean net payoff was not significantly different across the two conditions (349.81 JPY, *sd* = 83.95 vs. 356.44 JPY, *sd* = 52.56 in the signalling vs. standing conditions, respectively), $t_{102} = 0.48$, $p = .630$; whereas in experiment 2, the mean net payoff was significantly smaller in the signalling condition (388.47 JPY, *sd* = 22.32) than in the standing condition (440.00 JPY, *sd* = 42.47), $t_{97} = 7.53$, $p < .001$. This result is understandable because, in the signalling condition, participants *wasted* some of their payoff in exchange for good standing. In contrast, in the standing condition, there was no option to waste their payoff. However, this does not

necessarily mean that the standing strategy is a more cost-effective strategy than *intSIG*. The standing strategy demands *less visible* cognitive cost. Notice that although the cognitive cost is less visible, it may have real consequences. If you are busy processing some information, you might overlook some fitness-relevant cues, such as cues of predators or nutrition-rich foods. Therefore, whether *intSIG* is an adaptive strategy might depend on the trade-off between the tangible signalling cost and the less-visible cognitive cost.

(d) Post-experiment questionnaire

In order to assess participants' strategies in more detail, we had participants fill out a vignette post-experiment questionnaire. In the questionnaire, we presented participants with every possible situation as a donor. In the signalling condition, they were presented the following three situations: the partner's choice in the previous round was 'gave', 'did not give + abandoned, and 'did not give + did not abandon'. In the standing condition, participants were presented the following four situations: GG, GN, NG, and NN. Given each of these situations, participants rated their impression of the recipient on a five-point scale (1 = 'very bad' to 5 = 'very good'), inferred the goodness of the recipient's intention in the previous round (1 = 'very bad' to 5 = 'very good'), and indicated how they would behave towards the recipient (either 'give' or 'not give'). In the signalling condition, we included two additional questions. When participants chose 'not give' to the third question, they were further asked whether to abandon their resource. Participants were also asked whether they would abandon their resource if an implementation error occurred.

The analyses of the responses to the post-experiment questionnaire paralleled the results reported in the main text. As shown in figure S1, in the signalling condition, participants' impression of the partner was influenced by the recipient's previous behaviour: $F_{2, 102} = 90.66$, $p < .001$ and $F_{2, 96} = 150.96$, $p < .001$ for experiments 1 and 2, respectively. A post-hoc test by Ryan's method indicated that participants' impression of the 'giver' was the most favourable in both experiments 1 and 2. More importantly, participants' impression of the 'signalling non-giver' was more favourable than

that of the ‘non-signalling non-giver’ in both experiments 1 and 2.

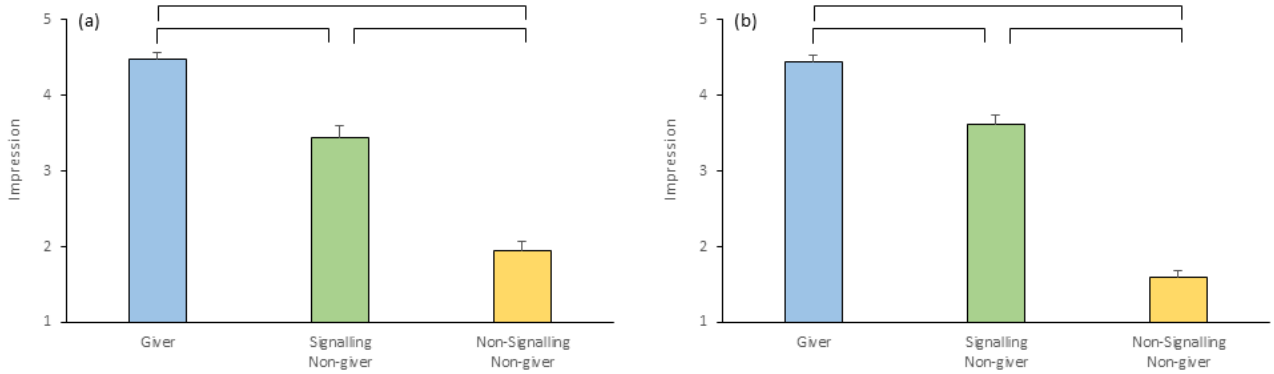


figure S1. Mean impression score as a function of the recipient’s previous behaviour in the signalling condition of (a) experiment 1 and (b) experiment 2. The error bars indicate standard error of the mean.

As for the inferred intention, as shown in figure S2, the effect of recipient type was significant: $F_{2, 102} = 53.46$, $p < .001$ and $F_{2, 96} = 49.40$, $p < .001$ for experiments 1 and 2, respectively. Again, participants attributed a more benign intent to the ‘giver’ and ‘signalling non-giver’ than to the ‘non-signalling non-giver’.

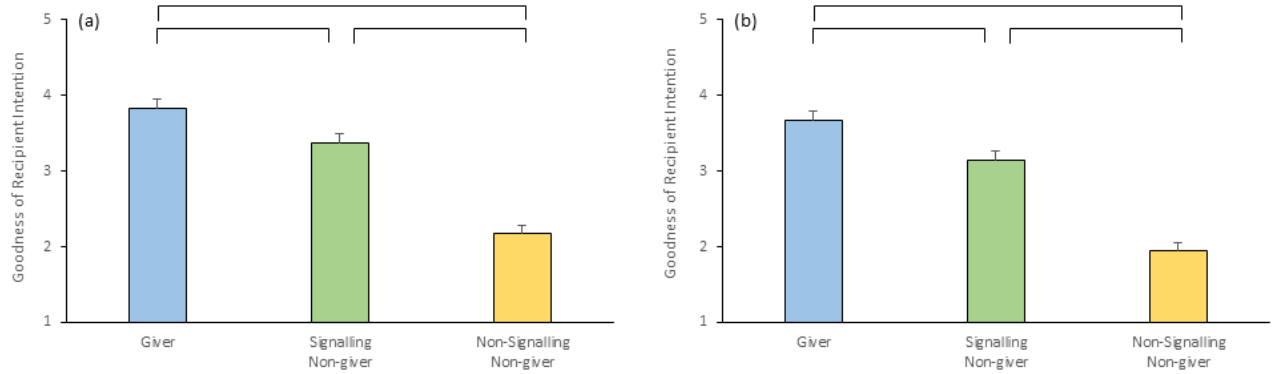


figure S2. Mean good-intention score as a function of the recipient’s previous behaviour in the signalling condition of (a) experiment 1 and (b) experiment 2. The error bars indicate standard errors of the mean.

Participants' hypothetical behaviour towards the recipient showed a similar pattern as their actual behaviour in the game experiment (see figure S3). We conducted a series of McNemar tests using the Bonferroni correction. The proportion of participants who chose 'give' was greater when the recipient was a 'giver' than when the recipient was a 'non-signalling non-giver' in both experiments 1 and 2 ($p < .001$ for each comparisons). More importantly, the proportion of participants who chose 'give' was greater when the recipient was a 'signalling non-giver' than a 'non-signalling non-giver' ($p < .001$ for each comparisons) in both experiments 1 and 2. Therefore, it was shown that the signal option was effective to amend the recipient impression, communicate the recipient's benign intent, and induce helping behaviour from future partners.

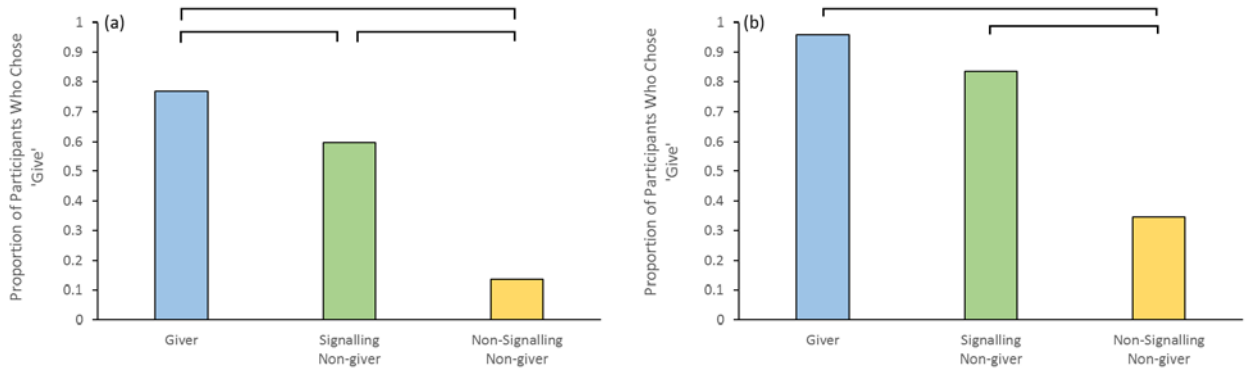


figure S3. Proportion of participants who chose 'give' as a function of recipient type in the signalling condition of (a) experiment 1 and (b) experiment 2.

In the signalling condition, we further assessed participants' willingness to use the signal option if they chose the 'not give' option in response to the previous question. As can be seen in figure S3, only a small portion of participants chose the 'not give' option in response to the 'giver' and 'signalling non-giver'. Therefore, it was impossible to test whether justified defectors, who chose 'not give' only to the 'non-signalling non-giver', are more likely to use the signal option than genuine defectors, who chose 'not give' to all three types of recipients. Accordingly, we only report the omnibus signal use rates here. Among those who chose 'not give' to the 'non-signalling

non-giver', 29% and 90% of participants (in experiments 1 and 2, respectively) reported willingness to use the signal option.

We also assessed participants' willingness to use the signal option after implementation error. The proportions of participants (experiments 1 and 2, respectively) who reported they would use the signal option at least once in the three situations (where the recipient was 'giver', 'signalling non-giver', and 'non-signalling non-giver', respectively) were 50% and 82% in experiment 1 and 2, respectively. There were 12 and one genuine defectors, who had never chosen 'give', and thus were not expected to experience implementation error. Once these participants were discarded, the proportions of signal users increased to 65% and 83%. In addition, 42% and 76% of participants (experiments 1 and 2) reported to use the signal option consistently across the three situations (the proportions increased to 55% and 77% once the genuine defectors were discarded from the data).

In the standing condition, we presented the four recipients (GG, GN, NG, and NN) to participants and asked the three following questions: impression of the recipient, perceived good intention of the recipient, and willingness to give. The results are mixed in terms of participants' discrimination of the NG and NN recipients. The main effect of the recipient type on impression score was significant, $F_{3, 153} = 94.87$, $p < .001$ and $F_{3, 147} = 111.46$, $p < .001$, in experiments 1 and 2. As shown in figure S4, post-hoc tests revealed that participants had a more favourable impression of GG and GN recipients than NG and NN recipients. Moreover, in both experiments, participants formed a better impression of NN than that of NG.

Participants attributed different levels of good intention to different types of recipients, $F_{3, 153} = 41.49$, $p < .001$ and $F_{3, 147} = 58.27$, $p < .001$ in experiments 1 and 2 (figure S5). Again, participants attributed a more benign intent to GG and GN recipients than to NG and NN recipients in both experiments 1 and 2. In addition, participants attributed a more benign intent to the NN recipient than NG recipient.

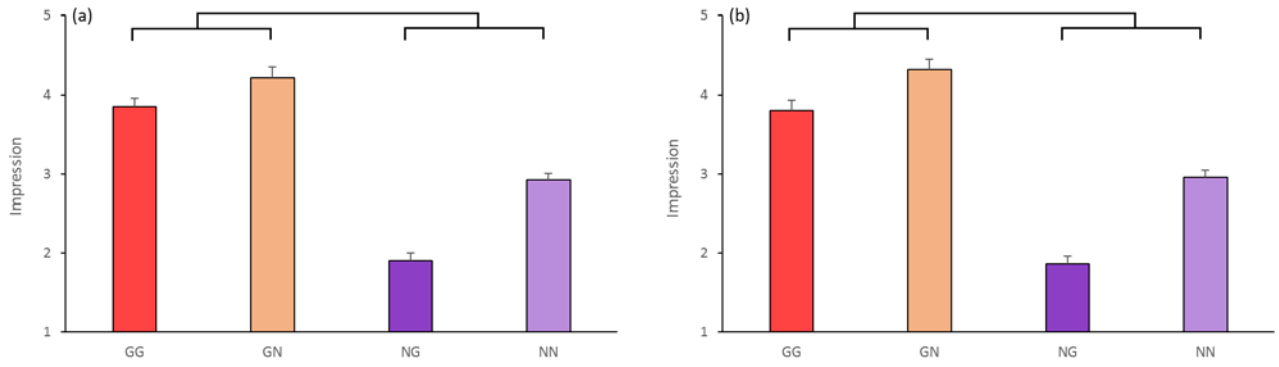


figure S4. Mean impression score as a function of recipient type (GG, GN, NG, and NN) in the standing condition of (a) experiment 1 and (b) experiment 2. The error bars indicate standard error of the mean.

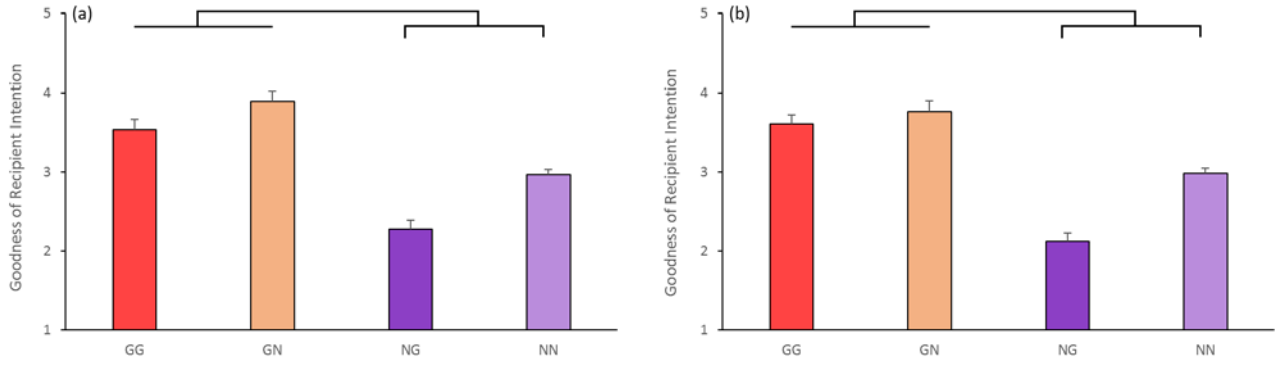


figure S5. Mean good-intention score as a function of recipient type (GG, GN, NG, and NN) in the standing condition of (a) experiment 1 and (b) experiment 2. The error bars indicate standard error of the mean.

As for willingness to give, we conducted a series of McNemar tests with the Bonferroni correction. The results were almost identical with the behavioural data reported in the main text (figure S6). Participants were more willing to give to GG and GN recipients than to NN and NG recipients ($p < .001$ for each comparisons). Moreover, participants did not differentiate NN and NG recipients ($p = 1.00$ and $p = .052$ in experiments 1 and 2, respectively). In sum, the results of the

post-experiment questionnaire provided mixed support for the standing strategy. Although participants perceived the NN recipient slightly more favourably than the NG recipient, they were not willing to treat the NN recipient more favourably than the NG recipient.

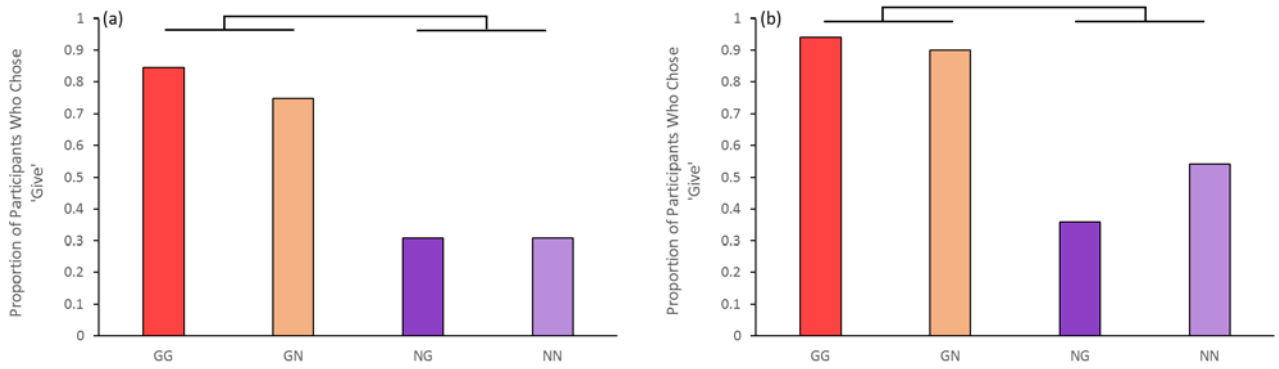


figure S6. Proportion of participants who chose 'give' as a function of recipient type (GG, GN, NG, and NN) in the standing condition of (a) experiment 1 and (b) experiment 2.

In addition to these questionnaire, in experiment 2, we asked participants to fill out the questionnaire containing the Japanese version of the Test of Self-Conscious Affect (TOSCA) [2], which was originally developed by Tangney and Dearing [3] to assess respondents' propensity to feel shame and guilt, along with some less focal emotions. Participants' trait shame and guilt scores were not related to the participants' behaviours in the donation game. Although we do not report the details of the results associated with the TOSCA here, interested readers can find the analysable data in the dataset attached to this article.

References

1. 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).
2. 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).
3. Tangney J P, Dearing R S. 2002 *Shame and guilt.* New York: Guilford Press.