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.

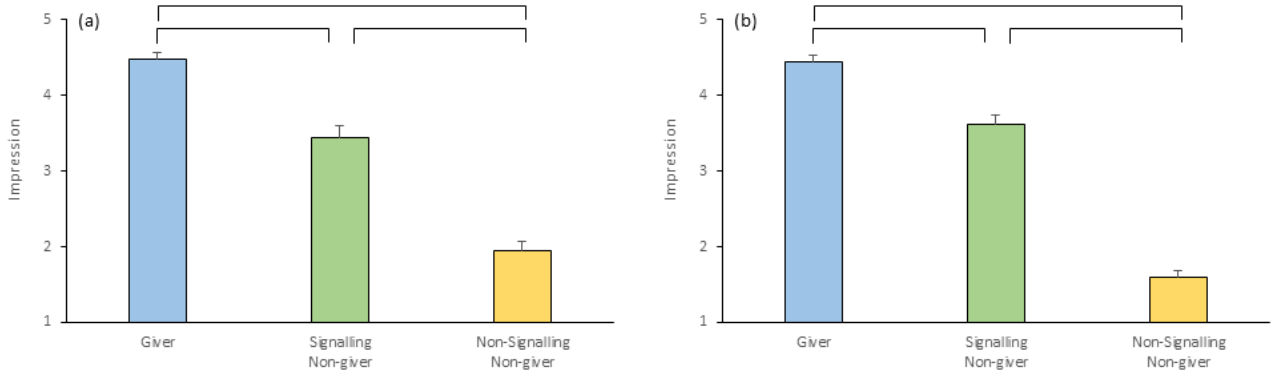

figure S1. Mean impression score as a function of the recipient’s previous behaviour in the signalling condition of (a) experiment 1 and (b) experiment 2. The error bars indicate standard error of the mean.

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’.

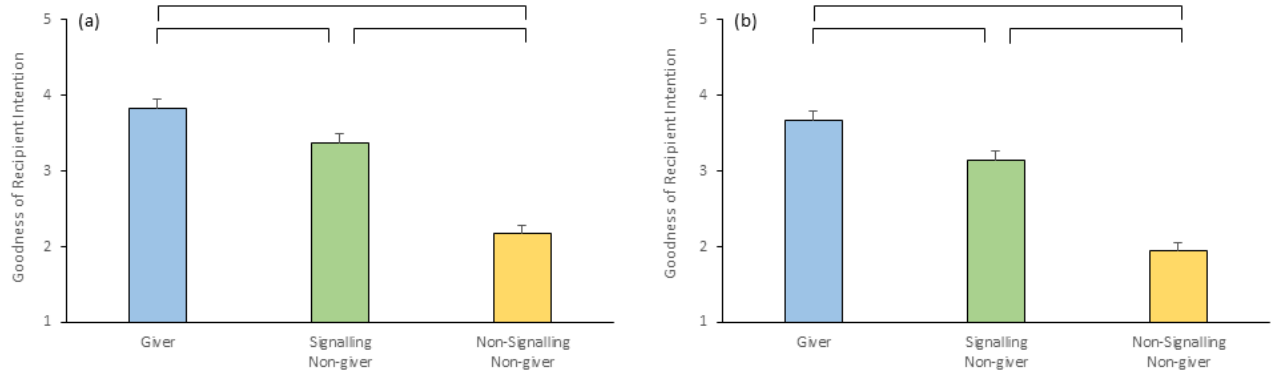

figure S2. Mean good-intention score as a function of the recipient’s previous behaviour in the signalling condition of (a) experiment 1 and (b) experiment 2. The error bars indicate standard errors of the mean.

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.

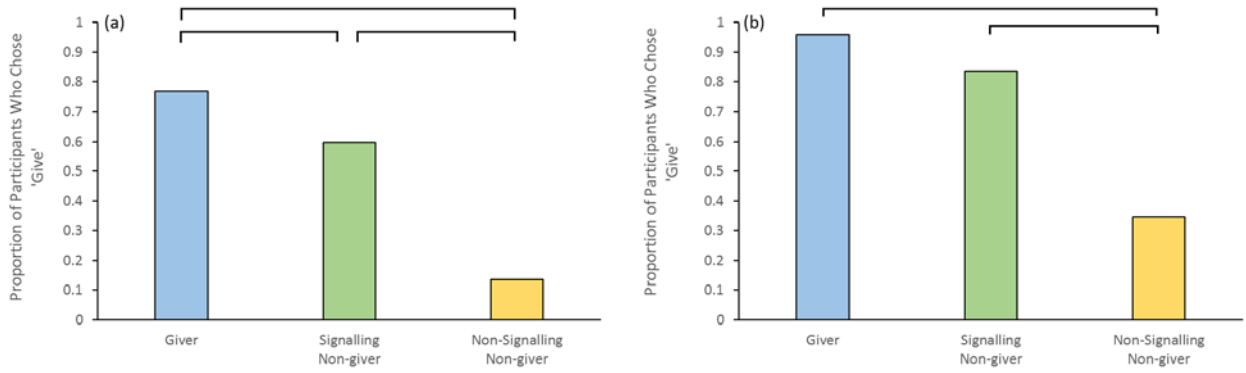

figure S3. Proportion of participants who chose 'give' as a function of recipient type in the signalling condition of (a) experiment 1 and (b) experiment 2.

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.

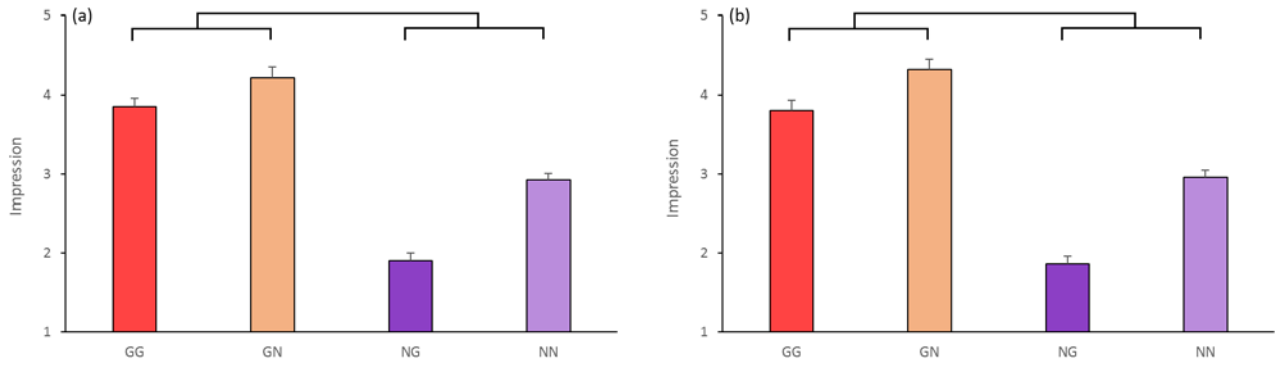

figure S4. Mean impression score as a function of recipient type (GG, GN, NG, and NN) in the standing condition of (a) experiment 1 and (b) experiment 2. The error bars indicate standard error of the mean.

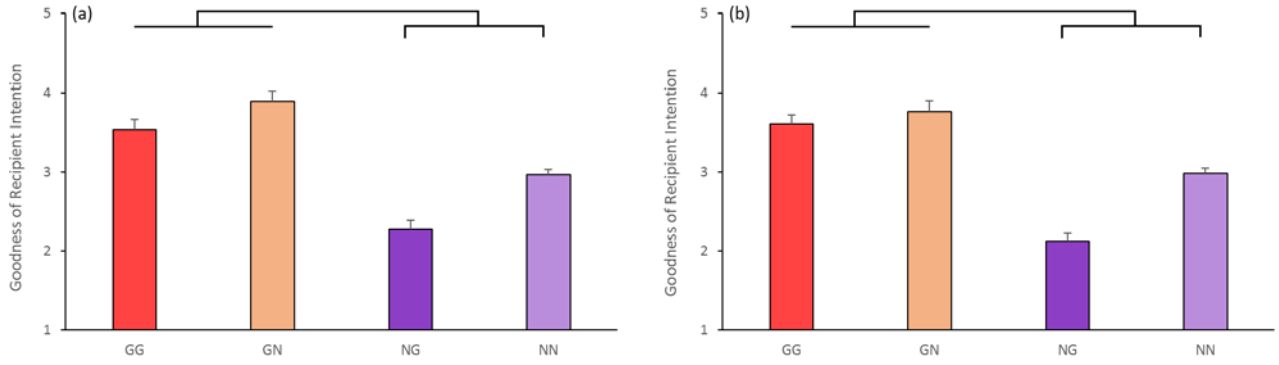

figure S5. Mean good-intention score as a function of recipient type (GG, GN, NG, and NN) in the standing condition of (a) experiment 1 and (b) experiment 2. The error bars indicate standard error of the mean.

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.

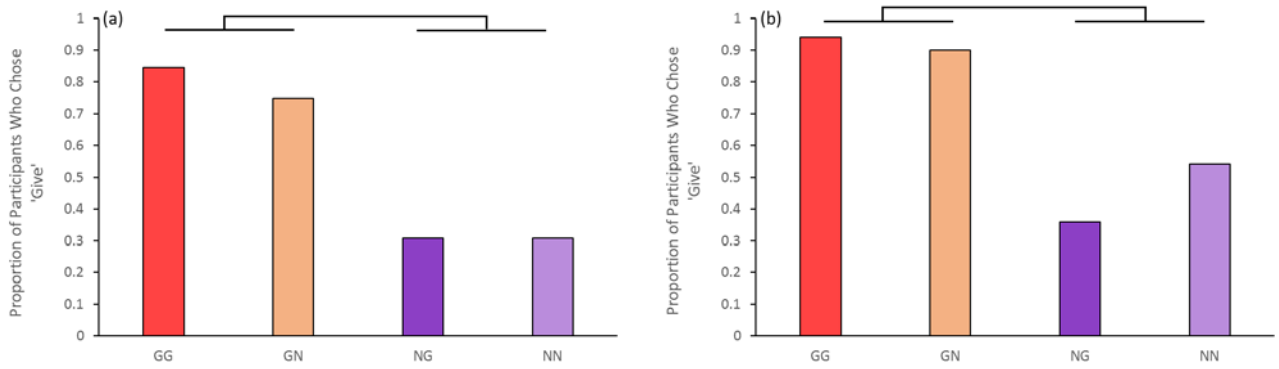

figure S6. Proportion of participants who chose 'give' as a function of recipient type (GG, GN, NG, and NN) in the standing condition of (a) experiment 1 and (b) experiment 2.

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.