
(This is a sample cover image for this issue. The actual cover is not yet available at this time.)

Journal of Theoretical Biology 277 (2011) 48–54

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

# Two wrongs don't make a right: The initial viability of different assessment rules in the evolution of indirect reciprocity

Karthik Panchanathan

*Center for Behavior, Evolution, and Culture, Department of Anthropology, University of California, Los Angeles, Haines Hall 341, Box 951553, Los Angeles, California 90095, USA*

ARTICLE INFO

*Article history:*
Received 23 September 2010
Received in revised form
8 February 2011
Accepted 9 February 2011
Available online 16 February 2011

*Keywords:*
Indirect reciprocity
Cooperation
Morality

ABSTRACT

Indirect reciprocity models are meant to correspond to simple moral systems, in which individuals assess the interactions of third parties in order to condition their cooperative behavior. Despite the staggering number of possible assessment rules in even the simplest of these models, previous research suggests that only a handful are evolutionarily stable against invasion by free riders. These successful assessment rules fall into two categories, one which positively judges miscreants when they refuse to help other miscreants, the other which does not. Previous research has not, however, demonstrated that all of these rules can invade an asocial population—a requirement for a complete theory of social evolution. Here, I present a general analytical model of indirect reciprocity and show that the class of assessment rules which positively judges a refusal to help scofflaws cannot invade a population of defectors, whereas the other class can. When rare, assessment rules which positively judge a refusal to help bad people produce a poor correlation between reputation and behavior. It is this correlation that generates the assortment crucial in sustaining cooperation through indirect reciprocity. Only assessment rules that require good deeds to achieve a good reputation guarantee a strong correlation between behavior and reputation.



## 1. Introduction

Indirect reciprocity models capture the idea that people assess the interactions of third parties in order to condition their cooperative behavior (Nowak and Sigmund, 1998, 2005). In direct reciprocity, individuals condition their behaviors on their partners' previous behaviors; in indirect reciprocity, individuals condition their behaviors on their partners' reputations, which are summary representations of past interactions with third parties. Indirect reciprocity might, therefore, be construed as a simple moral system in which individuals assess the actions of others and assign to them reputations.

Cooperation via indirect reciprocity can evolve when reputation correlates with past behavior (Nowak and Sigmund, 1998; Leimar and Hammerstein, 2001; Panchanathan and Boyd, 2003). This correlation allows help to be channeled to those who have been helpful and withheld from those who have not. Such assortment is the key to understanding the evolution of reciprocity. In direct reciprocity, behavioral assortment is generated within dyads (Axelrod and Hamilton, 1981); in indirect reciprocity, assortment is generated across dyads, such that communities enforce norms of conduct (Kandori, 1992; Nowak and Sigmund, 2005).

The logic of indirect reciprocity becomes more transparent when assessment rules (also called 'assessment modules' in Brandt and Sigmund, 2004, and 'reputation dynamics' in Ohtsuki and Iwasa, 2004) are disentangled from behavioral strategies (also called 'action modules' in Brandt and Sigmund, 2004, and 'behavioral strategies' in Ohtsuki and Iwasa, 2004). In a social exchange in which a donor can help a recipient, an assessment rule assigns to the donor a score (in the most simple case, 'good' or 'bad') based on inputs such as the donor's prior score, the recipient's prior score, and whether or not help was provided. A behavioral strategy determines whether or not the donor will help based on the recipient's score (and sometimes the donor's score, too).

Searching 4096 combinations of assessment rules and behavioral strategies (all the possible combinations in a relatively simple model of indirect reciprocity), Ohtsuki and Iwasa (2004) found eight assessment rule—behavioral strategy combinations which, when common, maintain high levels of cooperation, resisting invasion by mutants who never help. The assessment rules on this list have three common features: helping a good recipient is considered a good deed, refusing to help a good recipient is bad, and refusing to help a bad recipient does not incur a penalty. In this paper, I focus on one of the dimensions on which these eight assessment

*E-mail address:* karthikpanchanathan@gmail.com

0022-5193/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2011.02.009

rule—behavioral strategy combinations differ: how should a community judge a bad person who refuses to help another bad person? On the one hand, the refusal can be construed as a form of norm enforcement (e.g., 'do not help bad people'), and, therefore, be considered a good thing. On the other hand, merely refusing to help bad people costs nothing, and, therefore, may not deserve redemption. While intuition may suggest one alternative or the other as the only sensible solution, the logic of Ohtsuki and Iwasa's model is sound; cooperation based on indirect reciprocity can be evolutionarily stable under either assessment rule.

In order to make the distinction between these alternative assessment norms more intuitive, I introduce the following terms: *two wrongs* refers to an assessment rule which confers good standing on a bad donor who refuses to help a bad recipient; *unforgiven* refers to an assessment rule which does not confer good standing on a bad donor who refuses to help a bad recipient.

Based on previous research, it seems that the *two wrongs* rule outperforms the *unforgiven* rule when the two are pitted against each other (Takahashi and Mashima, 2006; Pacheco et al., 2006; Chalub et al., 2006; Scheuring, 2009). This research used agent-based simulations, rather than analytical techniques. As such, it is not clear why one rule does better than the other. In these particular agent-based simulations, the researchers included perception errors; in models using analytical methods, perception errors are not considered. With an analytical model, the researcher must assume that all members of a community agree on the reputations of each other. With an agent-based simulation, individuals can have private representations of fellow community members, which need not correlate with the representations of others. Unless there are real-world mechanisms which generate consensus opinions, the realism afforded by agent-based simulations seems desirable (see Nakamaru and Kawata, 2004 for a model of gossip), a point to which I will return in the discussion.

While previous research has compared different assessment rules in terms of evolutionary stability, finding that indirect reciprocity can be based on either the *two wrongs* or the *unforgiven* rule, little emphasis has been given to initial viability. An explanation for the evolution of cooperation must presuppose an asocial ancestral state. Without demonstrating that cooperative strategies can increase when rare, our theories are necessarily incomplete (Axelrod and Hamilton, 1981). In this paper, I develop a general model of indirect reciprocity to demonstrate that the *two wrongs* and the *unforgiven* assessment rules have quite different invasion properties. The *two wrongs* assessment rule, which confers good standing on bad donors who do not help bad recipients, is unlikely to increase when rare against a resident population of defectors, unless the benefit-cost ratio of social exchange is high. The alternate rule, *unforgiven*, which does not reward bad donors who refuse to help bad recipients, can increase when rare even when the gains from social exchange are meager.

The reason for this difference has to do with how well reputation predicts behavioral strategy under each rule. Assortment is the key to the evolution of cooperation. In indirect reciprocity, assortment is generated through reputation. Cooperation is viable when there is a strong correlation between having a 'good' reputation and having a cooperative strategy. Under the *two wrongs* rule, the correlation between reputation and behavioral strategy is a function of the frequency of reciprocators. When reciprocators following the *two wrongs* rule are common, there is a strong correlation between reputation and behavioral strategy and so cooperation is evolutionarily stable. When reciprocators using the *two wrongs* rule are rare, however, the correlation weakens, and so invasion becomes unlikely. Under the *unforgiven* rule, the correlation between reputation and behavioral strategy does not depend on the frequency of reciprocators.

To get an intuitive understanding of why the effectiveness of the *two wrongs* assessment rule varies as a function of its frequency, consider the following two scenarios. When the population is comprised mostly of reciprocators using the *two wrongs* rule, unless there are errors in strategy execution, individuals will see their fellow community members as good and cooperate with them. A mutant defector will refuse to help his partners. Because his partners are good, the defector will be seen as bad for having refused to help them. In this scenario, reputation accurately predicts behavioral strategy: reciprocators are good, defectors are bad. Now, suppose that the population is comprised mostly of defectors. These individuals never help one another. A mutant reciprocator, observing a bad donor refuse to help a bad recipient, will think positively of the donor. Because many of the observed interactions will be like this, reputation does not accurately predict behavioral strategy: defectors are often labeled as good. The *two wrongs* assessment rule closely mirrors the *PAVLOV* strategy in models of direct reciprocity (Nowak and Sigmund, 1993; Boerlijst et al., 1997).

As previously mentioned, analytic models of indirect reciprocity must ignore perception errors, seemingly an unreasonably optimistic assumption. Agent-based simulations of indirect reciprocity suggest that selection favors the *two wrongs* rule over the *unforgiven* rule (Takahashi and Mashima, 2006; Pacheco et al., 2006; Chalub et al., 2006; Scheuring, 2009). As with the literature on direct reciprocity (Boerlijst et al., 1997), it might be the case that assessment rules like *unforgiven* get cooperation off the ground, only to be replaced by rules like *two wrongs*.

There is a good reason why the *two wrongs* rule might do well in a world with perception errors. Perception errors can destroy trust within a community. Otherwise cooperative individuals may come to believe their fellow community members to be cheats. Under a rule like *contrite tit for tat* in direct reciprocity (Boyd, 1989) or *unforgiven* in indirect reciprocity, there is no way out this stalemate. If everyone believes everyone else to be bad, no one will cooperate. If no one cooperates, no one can regain good standing. Rules like *PAVLOV* in the context of direct reciprocity (Nowak and Sigmund, 1993; Boerlijst et al., 1997) and *two wrongs* in the context of indirect reciprocity can break out of this vicious cycle. If everyone believes everyone else to be bad, no one will cooperate. Observers look favorably on the refusal to help bad-standing partners, thereby resetting cooperation.

## 2. The model

I assume an infinite and unstructured population. Individuals are paired up with a randomly selected partner. Each individual acts as a donor, having an opportunity to help his partner, the recipient, providing her a benefit $b$, at a cost $-c$ to himself. Both partners act as donors and recipients in the same interaction. Because decisions are simultaneous, individuals cannot condition their behavior as donors on their partners' behavior as donors. Pairs are then broken up. With probability $w$, new pairs are formed and another round of social interaction ensues; with probability $1-w$, social interaction ends and individuals reproduce at a rate proportional to their fitness. On average, there will be $1/(1-w)$ rounds of social interaction. Because the population is infinite, individuals never pair up with the same partner twice, precluding direct reciprocity (Axelrod and Hamilton, 1981).

To keep the model simple without loss of generality, I assume that individuals never make a mistake when deciding whether or not to help. Thus, individuals who intend to help, help; individuals who intend not to help, do not help.

After each round of social interaction, individuals observe the behavior of others in the population and assign to each a reputation,

which can be either good or bad, *G* or *B*. All individuals in the population share the same assessment rule (see Uchida and Sigmund, 2010, and Pacheco et al., 2006, for models which allow different assessment rules within the same population to compete with one another). Reputation broadcast is complete such that individuals know the reputation of everyone else in the population. Again, this assumption is made for simplicity; it does not alter the qualitative results. Finally, I assume no perception errors. Reputations accurately reflect past behavior as interpreted by the community's assessment rule. While obviously unrealistic, this assumption allows analytic tractability. In the discussion, this assumption is justified and discussed in the context of indirect reciprocity models.

In models of indirect reciprocity, individuals are characterized by both a behavioral strategy and an assessment rule (Brandt and Sigmund, 2004; Ohtsuki and Iwasa, 2004). I consider two behavioral strategies, *reciprocator* and *defector*. A *reciprocator* helps good recipients, paying the cost $-c$, and withholds help from bad recipients. This behavioral strategy does not address the first round of social interaction; because no previous interactions have taken place, no one has a reputation. I assume that *reciprocators* help their partners in the first round. Assuming cooperation evolves, selection should favor such a rule in the first round in models of indirect reciprocity (Panchanathan and Boyd, 2003). *Defector*, as the name implies, never helps. Let $p$ denote the fraction of the population comprised of *reciprocators* and $1-p$ the fraction playing *defector* (see Table 1 for a complete list of the model parameters).

An assessment rule is a function. It takes as input three pieces of information: the reputation of the donor prior to the social interaction (*G* or *B*), the reputation of the recipient prior to the interaction (*G* or *B*), and whether or not the donor helped the recipient (*H* or *N*). Based on these inputs, an assessment rule outputs the new reputation for the donor (*G* or *B*). There are $2^8$ possible assessment rules. Ohtsuki and Iwasa (2004) found eight of these assessment rules to be evolutionarily stable when combined with a reciprocating behavioral strategy (listed in Table 2).

These eight assessment rules have the following common features:

- Helping a good partner results in a good reputation (columns *GGH* and *BGH*).
- Not helping a good partner results in a bad reputation (columns *GGN* and *BGN*).
- A good person retains his good reputation if he refuses to help a bad partner (column *GBN*).

**Table 1**
Model parameters.

| Parameter | Definition |
|---|---|
| $b$ | Benefit of receiving help |
| $c$ | Cost of helping |
| $w$ | Probability of another social interaction |
| $p$ | Fraction of the population playing the *reciprocator* strategy |
| $r$ | Exogenous assortment (e.g., kin-biased interactions) |
| $\hat{g}$ | Equilibrium fraction of the population with a good reputation |
| $\hat{g}_R$ | Equilibrium fraction of *reciprocators* with a good reputation |
| $\hat{g}_D$ | Equilibrium fraction of *defectors* with a good reputation |
| $\pi_R$ | Fitness of *reciprocators* |
| $\pi_D$ | Fitness of *defectors* |
| $\pi_0$ | Baseline fitness |
| $P\{R\|R\}$ | Probability a *reciprocator* interacts with a *reciprocator* |
| $P\{D\|R\}$ | Probability a *reciprocator* interacts with a *defector* |
| $P\{R\|D\}$ | probability a *defector* interacts with a *reciprocator* |
| $P\{D\|D\}$ | Probability a *defector* interacts with a *defector* |

**Table 2**
Ohtsuki and Iwasa's (2004) 'Leading 8' assessment rules in a model of indirect reciprocity. The assessment rules are listed in the left-most column, numbered 1–8. The next eight columns represent the eight different types of social interaction in this model. Each column header is a three letter code (e.g., *GGH*). The first letter denotes the reputation of the donor prior to the social interaction; the second letter denotes the prior reputation of the recipient; and the third letter indicates whether or not the donor helped the recipient (*H* or *N*). For each assessment rule, there is a row of eight letters representing the donor's reputation (*G* or *B*) as a result of each interaction.

| Rule | GGH | GGN | GBH | GBN | BGH | BGN | BBH | BBN |
|---|---|---|---|---|---|---|---|---|
| 1 | G | B | G | G | G | B | G | B |
| 2 | G | B | G | G | G | B | B | B |
| 3 | G | B | B | G | G | B | G | B |
| 4 | G | B | B | G | G | B | B | B |
| 5 | G | B | G | G | G | B | G | G |
| 6 | G | B | G | G | G | B | B | G |
| 7 | G | B | B | G | G | B | G | G |
| 8 | G | B | B | G | G | B | B | G |

**Table 3**
Assessment rule classes considered in this model. The two classes of assessment rule are listed in the left-most column, *unforgiven* and *two wrongs*. The next eight columns represent the eight different types of social interaction in this model. Each column header is a three letter code (e.g., *GGH*). The first letter denotes the reputation of the donor prior to the social interaction; the second letter denotes the prior reputation of the recipient; and the third letter indicates whether or not the donor helped the recipient (*H* or *N*). For each assessment rule, there is a row of eight letters representing the donor's reputation (*G* or *B*) as a result of each interaction. Because the behavioral strategies considered in this model do not help recipients of bad-standing (and never mistakenly do so), columns *GBH* and *BBH* are impossible social interactions, denoted by asterisks.

| Rule | GGH | GGN | GBH | GBN | BGH | BGN | BBH | BBN |
|---|---|---|---|---|---|---|---|---|
| *unforgiven* | G | B | * | G | G | B | * | B |
| *two wrongs* | G | B | * | G | G | B | * | G |

Because individuals always do what they intend (i.e., there are no strategy execution errors), individuals playing with either a *reciprocator* or *defector* strategy will never help a partner with a bad reputation. The qualitative results of this model do not change if strategy execution errors are included. To keep the model simple, I exclude these kinds of errors. So, two columns (*GBH* and *BBH*) can be eliminated from Table 2. Putting asterisks in for these impossible outcomes results in Table 3.

These two classes of assessment rule, *two wrongs* and *unforgiven*, differ only in how they view a bad donor refusing to help a bad recipient (column *BBN*). The *two wrongs* rule looks favorably on a bad donor who refuses to help a bad partner, while the *unforgiven* rule does not.

## 3. Reputation dynamics

We begin by writing equations for the reputation dynamics. By assumption, everyone starts out with a good reputation. In subsequent rounds, each individual's reputation is revised based on the situational context (comprising the prior reputation of both donor and recipient), the donor's decision to help or not (as determined by the donor's behavioral strategy), and the assessment rule of the population (in this case, either *two wrongs* or *unforgiven*).

For the *unforgiven* rule, the analysis is straightforward. After the first round of social interaction, all *defectors* fall into disrepute for having refused to help their partners, who were all good. As a good reputation can only be regained through an act of help, *defectors* will forever remain bad. *Reciprocators* cooperate in the

first round, meeting good partners, and so enter the second round with a good reputation. Thereafter, they either meet a good partner playing the *reciprocator* strategy, in which case they help, or they meet a bad partner playing the *defector* strategy, in which case they do not. In either case, *reciprocators* retain their good reputation. So, without strategy execution errors, *reciprocators* remain in good standing throughout their lives. At equilibrium, *reciprocators* are all good and *defectors* are all bad, denoted by $\hat{g}_R = 100\%$ and $\hat{g}_D = 0\%$.

As with the *unforgiven* rule, *reciprocators* will always have a good reputation under *two wrongs* rule. Without strategy execution errors, *reciprocators* will not defect on good partners and so never tarnish their reputations, implying $\hat{g}_R = 100\%$. The reputations of *defectors*, who never help, are determined by the reputations of their interaction partners. When a *defector* refuses to help a good recipient (columns *GGN* or *BGN*, Table 3), he earns a bad reputation; when he refuses to help a bad recipient (columns *GBN* or *BBN*, Table 3), he earns a good reputation, implying $\hat{g}_D = 1-\hat{g}$, where $\hat{g}$ denotes the fraction of the population with a good reputation at equilibrium, regardless of behavioral strategy. Letting $P\{D|D\}$ denote the probability of a *Defector* interacting with another *Defector*, we have

$$\hat{g}_D = \frac{P\{D|D\}}{1+P\{D|D\}} \tag{1}$$

Eq. (1) captures the verbal description above: the equilibrium fraction of *defectors* with a good reputation is a function of interaction probabilities. The more often *defectors* interact with other *defectors*, the higher this fraction will be, reaching a maximum of 50% when $P\{D|D\} = 1$.

## 4. Fitness functions

Let $\pi_R$ and $\pi_D$ denote the fitnesses of *reciprocators* and *defectors*, and $\pi_0$ the baseline fitness. And, let $P\{i|j\}$ be the probability of an individual playing the behavioral strategy $j$ interacting with a partner playing the behavioral strategy $i$.

$$\pi_R = P\{R|R\}b - c + \frac{w}{1-w}\left[P\{R|R\}\hat{g}_R(b-c) - P\{D|R\}\hat{g}_D c\right] + \pi_0 \tag{2}$$

$$\pi_D = P\{R|D\}b + \frac{w}{1-w}P\{R|D\}\hat{g}_D b + \pi_0 \tag{3}$$

In the first round, *reciprocators* help all partners, always paying the cost $-c$. *Reciprocators* and *defectors* receive help ($b$) when they are paired with a *reciprocator*, given by $P\{R|R\}$ and $P\{R|D\}$.

Following Panchanathan and Boyd (2003), in deriving fitness functions (2) and (3), I assume that $\hat{g}_R$ and $\hat{g}_D$ reach equilibrium in the second round. So long as $w$ is sufficiently close to 1, without which reciprocity could not evolve, little is lost with this approximation. This assumption merely implies that the timescale of reputation dynamics is fast relative to lifespan.

In rounds after the first, of which there will be $w/(1-w)$ on average, *reciprocators* help good partners and receive help from *reciprocators* when they themselves are good, while *defectors* receive help from *reciprocators* when they are good.

## 5. Evolutionary stability

Let us first analyze when *reciprocators* are evolutionarily stable against invasion by *defectors* under either assessment rule. Following Maynard Smith (1982), I assume *reciprocators* are common, which implies the following interaction probabilities: $Pr\{R|R\} \approx 1$, $Pr\{D|R\} \approx 0$, $Pr\{R|D\} \approx 1$, and $Pr\{D|D\} \approx 0$.

Substituting these interaction probabilities into fitness functions (2) and (3), and solving for $\pi_R > \pi_D$, results in the following stability condition for *reciprocators*:

$$w > \frac{c}{b(\hat{g}_R - \hat{g}_D) + c(1-\hat{g}_R)} \tag{4}$$

In reciprocity models, cooperation can be evolutionarily stable when the future casts a long shadow (i.e., $w$ is close to 1). Inequality (4) shows the same situation in a general indirect reciprocity model. As $\hat{g}_R$ gets close to 100% and $\hat{g}_D$ gets close to 0%, the conditions for cooperation based on reciprocity are most ripe. The difference between $\hat{g}_R$ and $\hat{g}_D$ represents the degree of assortment generated by reputation. When $\hat{g}_R - \hat{g}_D = 100\%$, assortment is complete: all *reciprocators* are considered *good* and all *defectors* are considered bad, which means that help is channeled exclusively to *reciprocators*. As the difference between $\hat{g}_R$ and $\hat{g}_D$ becomes smaller, assortment worsens; more and more help is given to *defectors*.

With either the *unforgiven* or *two wrongs* assessment rule, *reciprocators* always maintain a good reputation ($\hat{g}_R = 100\%$). With the *unforgiven* rule, *defectors* will always be bad ($\hat{g}_D = 0\%$). With the *two wrongs* rule, the fraction of *defectors* with a good reputation is governed by Eq. (1). With $P\{D|D\} \approx 0$, all *defectors* will be bad ($\hat{g}_D = 0\%$). When *reciprocators* are common, the equilibrium fractions of good *reciprocators* and good *defectors* are the same under either assessment rule. Inequality (4) reduces to $w > c/b$, which is the same stability condition Axelrod and Hamilton (1981) found for *tit for tat* in their model of direct reciprocity. This should not be surprising. When reputation perfectly correlates with behavioral strategy (all *reciprocators* are good and all *defectors* are bad), the dynamics of direct and indirect reciprocity are identical. In direct reciprocity, individuals condition their decision to help on the previous helping behavior of their partner; in indirect reciprocity, individuals condition their decision to help on how helpful their partners were to third parties.

## 6. Initial viability

When the future casts a long shadow (i.e., Inequality (4) is satisfied), cooperation based on indirect reciprocity is evolutionarily stable with either the *unforgiven* or *two wrongs* assessment rule. However, the *defector* strategy is also evolutionarily stable. This is true in all models of reciprocity. If we want to explain the evolution of cooperation, and assume that the ancestral condition was uncooperative, we need to explain how *reciprocators* can invade a population of *defectors*.

Axelrod and Hamilton (1981) showed that cooperation based on reciprocity could invade an asocial population when there is some assortment above and beyond that generated through reciprocity. This exogenous assortment might reflect, for example, kin-biased interactions. The synergy between kinship and reciprocity can drive social evolution. To see why, consider two extremes. Without any exogenous assortment, reciprocators mostly interact with defectors. They are taken advantage of in the first round, and thereafter refuse to cooperate. This first-round deficit prevents reciprocators from invading. Suppose instead that exogenous assortment is complete, reciprocators only interact with reciprocators, defectors only with defectors. The exogenous assortment has effectively segmented the population into two types of interaction pairs, reciprocating ones and defecting ones. None of the cooperation leaks out of the reciprocating pairs. So, reciprocators do better than defectors. Defectors are actually hurt by the exogenous assortment as they are more likely to meet other defectors.

We can ask a similar question for the evolution of cooperation based on indirect reciprocity: With some exogenous assortment,

can *reciprocators* invade a population of *defectors*? Because *defectors* are common, their fitness will be dominated by interactions with other *defectors*. For *reciprocators*, we assume some exogenous assortment, such that a mutant *reciprocator* will meet a fellow *reciprocator* with probability $r$, otherwise interacting with a randomly selected individual, which in this case will invariably be a *defector*. We can formalize assortment with the following interaction probabilities:

$$Pr\{R|R\} = r \tag{5}$$

$$Pr\{D|R\} = 1-r \tag{6}$$

$$Pr\{R|D\} = 0 \tag{7}$$

$$Pr\{D|D\} = 1 \tag{8}$$

Substituting these interaction probabilities into fitness functions (2) and (3) and solving for $\pi_R-\pi_D > 0$, we have the following invasion condition:

$$rb-c+\frac{w}{1-w}\left[rb\hat{g}_R-c[r\hat{g}_R+(1-r)\hat{g}_D]\right] > 0 \tag{9}$$

When $w=0$, inequality (9) reduces to $rb > c$ which is analogous to Hamilton's rule (Hamilton, 1964). When there is only one round of social interaction, there is no scope for reciprocity; cooperation can only evolve when the marginal benefit of help multiplied by the relatedness between donor and recipient is greater than the marginal cost.

If we assume that $rb < c$, the reciprocity term must be sufficiently large to offset the first round deficit in order for cooperation based on indirect reciprocity to increase when rare. Fitness from indirect reciprocity increases with the duration of social interaction ($w$), exogenous assortment ($r$), the net benefit of help ($b-c$), and, crucially, behavioral assortment generated by reputation ($\hat{g}_R-\hat{g}_D$).

For the *unforgiven* assessment rule, behavioral assortment generated through reputation is maximal ($\hat{g}_R = 100\%$ and $\hat{g}_D = 0\%$), so the invasion condition is

$$r > \frac{c(1-w)}{b-cw} \tag{10}$$

This is the standard result from reciprocity models (Axelrod and Hamilton, 1981; Panchanathan and Boyd, 2003). With small amounts of relatedness, cooperation based on indirect reciprocity can invade and dominate. To see this, we can take the limit of $r$ as $w \to 1$:

$$\lim_{w\to 1} r = 0 \tag{11}$$

For the *two wrongs* assessment rule, behavioral assortment is weaker ($\hat{g}_R = 100\%$ and $\hat{g}_D = 50\%$, from Eq. (1)), resulting in the invasion condition:

$$r = \frac{c(1-0.5w)}{b-0.5cw} \tag{12}$$

Taking the limit of $r$ as $w \to 1$, we have

$$\lim_{w\to 1} r = \frac{c}{2b-c} \tag{13}$$

The synergy between exogenous assortment and reputation-based assortment seen with the *unforgiven* rule is weaker with the *two wrongs* assessment rule. Unless the benefit-cost ratio of helping is high, the requisite exogenous assortment is unreasonably high for indirect reciprocity based on the *two wrongs* rule to evolve (Table 4).

The reason for the striking difference between the two assessment rules has to do with how each interprets a social interaction when a bad donor refuses to help a bad recipient (Table 3, final column). When *reciprocators* are rare, most of the interactions they observe will be between two *defectors*. As the *unforgiven* rule does not offer redemption for refusing to help a bad recipient, *defectors* will never be labeled as good. In contrast, the *two wrongs* rule considers this same refusal to help a bad recipient a good deed. Thus, the common-type *defector* will continuously flip between being good and bad, depending on whether he refuses to help a good or bad recipient. When *reciprocators* are rare, there is a poor correlation between behavioral strategy and reputation under the *two wrongs* assessment rule, and so the evolution of cooperation based on indirect reciprocity is unlikely. "Ohtsuki and Iwasa (2004, pp. 115) suggest the same: 'It is true that those four ESS pairs [the ones I have labeled *two wrongs*] are stable against a few ALLD-strategists, but they may be susceptible to a cluster of ALLD-strategists since defection between two cheaters is considered good under those rules.'"

**Table 4**
Exogenous assortment ($r$) required for indirect reciprocity to increase when rare. $N$ denotes the expected number of social interactions, which is given by $1/(1-w)$. The cost of help is assumed to be 1, while the benefit of receiving help varies.

| $N$ | *Unforgiven* $b=2$ | *Two wrongs* | | |
|---|---|---|---|---|
| | | $b=2$ | $b=4$ | $b=16$ |
| 1 | 0.50 | 0.50 | 0.25 | 0.06 |
| 2 | 0.33 | 0.43 | 0.20 | 0.05 |
| 4 | 0.20 | 0.39 | 0.17 | 0.04 |
| 16 | 0.06 | 0.35 | 0.15 | 0.03 |
| 256 | 0.004 | 0.33 | 0.14 | 0.03 |

## 7. Discussion

In this paper, I presented a simple model of indirect reciprocity in order to evaluate the initial viability of two types of assessment rule: *two wrongs* which assigns a good reputation when a bad donor refuses to help a bad recipient, and *unforgiven* which does not. While cooperation based on indirect reciprocity can be evolutionarily stable against mutants who never help under either assessment rule, the *two wrongs* rule is unlikely to invade when rare unless the benefit-cost ratio is high, whereas the *unforgiven* rule can invade even with a small, positive benefit-cost ratio.

The reason that the *two wrongs* rule does poorly has to do with how well reputation predicts behavioral strategy, the assortment undergirding cooperation via indirect reciprocity. A community using the *two wrongs* rule positively views a refusal to help a bad recipient, whatever the prior reputation of the donor. A good reputation can be gained without ever having to engage in costly help. When *reciprocators* are rare, most of the observed interactions will be between two *defectors*. Refusing to help one another, these *defectors* will often be in good standing, even though they never helped anyone! When *reciprocators* are common, a *defector* in the role of donor will likely interact with a *reciprocator* in good standing. The same refusal to help will no longer garner a good reputation. Under the *two wrongs* assessment rule, the likelihood that a *defector* is considered good depends on the frequency with which he interacts with other *defectors*. Assuming interactions are random with respect to behavioral strategy ($r = 0$), we can use Bayes' Theorem to compute the probability that an individual is a *reciprocator* given he has a good reputation under the *two wrongs* rule, which is given by: $P\{Reciprocator \mid \text{Good Standing}\} = p(2-p)$. As seen in Table 5, under the *two wrongs* assessment rule, reputation only becomes a good predictor of behavioral strategy when the fraction of *reciprocators* is close to one.

Previous researchers, using agent-based simulations, have found that the *two wrongs* assessment rule outperforms the

**Table 5**
Reputation as a predictor of behavioral strategy under the *two wrongs* assessment rule. $p$ denotes the fraction of the *reciprocators* in the population and $P\{Reciprocator|\text{Good Standing}\}$ denotes the probability that someone with a good reputation is a *reciprocator*. Interactions are assumed to be random with respect to behavioral strategy ($r=0$).

| $p$ | $P\{Reciprocator\|\text{Good Standing}\}$ |
|---|---|
| 0.00 | 0.00 |
| 0.25 | 0.44 |
| 0.50 | 0.75 |
| 0.75 | 0.94 |
| 1.00 | 1.00 |

*unforgiven* rule (Takahashi and Mashima, 2006; Pacheco et al., 2006; Chalub et al., 2006; Scheuring, 2009). One important difference between the model I presented here and previous work has to do with perception errors. Here, I assumed that everyone agrees on the reputations of each community member. In the previous work, individuals privately represent the reputations of one another, and so there is no guarantee of consensus.

To understand why perception errors are so pernicious, let us consider a community using the *unforgiven* assessment rule, which does not award good standing when a bad donor refuses to help a bad recipient. Assuming *reciprocators* are common, reputation is an effective assortment device, channeling cooperation to reciprocators, withholding it from defectors. Now, suppose an individual witnesses a social exchange between two others, misperceiving cooperation for defection. The observer now tags the donor as bad. If the two subsequently interact, the observer will refuse to help the previous donor. This refusal seems justified in his mind, but others, who did not misperceive the original exchange will not agree. This perception error will ripple through the community, leaving distrust in its wake, until each community member believes himself to be good and everyone else bad. Despite the fact that each member of the social group has the potential to cooperate, no one will because of mutual distrust.

The situation is different when a community uses the *two wrongs* rule. Under this rule, the community positively evaluates a bad donor who refuses to help a bad recipient. This feature of the *two wrongs* rule acts as a reset button, restoring trust to the community. A similar logic, in the context of dyads rather than a community, applies to the *PAVLOV* strategy in models of direct reciprocity (Nowak and Sigmund, 1993; Boerlijst et al., 1997).

While the maxim 'let bygones be bygones' may suggest we use something like *PAVLOV* at least some of the time in the context of dyadic reciprocity (perhaps only for deep friendships), it is not clear how to evaluate the *two wrongs* analog in the context of indirect reciprocity. While the rule can ameliorate the effects of perception errors, this happens not within dyads but across them, at the level of communities. Such a process does not seems to have intuitive appeal.

Ultimately, it is an empirical question what kinds of assessment rules govern reputation-based reciprocity in real-world human communities. More research is certainly needed on assessing the assessment rules we use. In Mashima and Takahashi (2008), subjects read scenarios in which targets either do or do not help recipients who are either good or bad. Subjects positively evaluated targets who withhold help from bad recipients. However, Mashima and Takahashi never explicitly stated whether the donors were good or bad community members. Further, in the vignettes, they wrote, 'All people usually help each other when they do farm work.' Subjects might have inferred the target to be an upstanding community member, and so the positive evaluation of his refusal to help does not necessarily reveal what subjects would think if they were told that the target was bad.

Empirics aside, from a theoretical perspective, it is important to consider what has been left out of published indirect reciprocity models with perception errors. Reciprocity models invariably preclude partner choice (Hammerstein, 2003; but see Enquist and Leimar, 1993). What would happen to a model of indirect reciprocity based on the *two wrongs* assessment rule if partner choice were permitted? *Reciprocators* might prefer to interact with good partners, hoping for a beneficial social exchange. *Defectors* might have a reputation-dependent preference. When good, a *defector* may attempt to pair up with a good partner, hoping to exploit a *reciprocator*. When bad, he may seek out a bad partner. By refusing to help his bad partner, the *defector* can regain a good reputation. In this way, defectors may undermine cooperation based on indirect reciprocity when the community uses the *two wrongs* assessment rule. The *unforgiven* rule would not be susceptible to strategic partner choice because good standing can only be gained through good deeds.

Indirect reciprocity models are also limited in assuming relatively unsophisticated information transfer mechanisms. The standard assumption is that individuals observe third-party interactions and privately form representations of what they witnessed. With language, however, assessments can be formed through a combination of observation and gossip. If perception errors are independent, then gossiping with others seems a powerful way of filtering signal from noise. Gossip might be especially effective if individuals use a conformist-bias (Boyd and Richerson, 1985) when assessing third parties. In a vignette study, Hess and Hagen (2006) find that subjects are more likely to believe gossip when it comes from multiple and/or independent sources. In an experimental economics game, Sommerfeld et al. (2008) find that subjects were more likely to cooperate with unknown partners when provided with more positive information about partners. With a more sophisticated gossip mechanism, the *unforgiven* rule may not be as susceptible to perception errors.

## Acknowledgments

I would like to thank Rob Boyd, H. Clark Barrett, Siamak Naficy, and the anonymous reviewer for reading and improving this paper.

## References

Axelrod, R., Hamilton, W.D., 1981. The evolution of cooperation. Science 211 (4489), 1390–1396.
Boerlijst, M.C., Nowak, M.A., Sigmund, K., 1997. The logic of contrition. Journal of Theoretical Biology 185, 281–293.
Boyd, R., 1989. Mistakes allow evolutionary stability in the repeated prisoner's dilemma game. Journal of Theoretical Biology 136, 47–56.
Boyd, R., Richerson, P.J., 1985. Culture and the Evolutionary Process. University of Chicago Press, Chicago.
Brandt, H., Sigmund, K., 2004. The logic of reprobation: assessment and action rules for indirect reciprocity. Journal of Theoretical Biology 231, 475–486.
Chalub, F.A., Santos, F.C., Pacheco, J.M., 2006. The evolution of norms. Journal of Theoretical Biology 241, 233–240.
Enquist, M., Leimar, O., 1993. The evolution of cooperation in mobile organisms. Animal Behaviour 45, 747–757.
Hamilton, W.D., 1964. The genetical evolution of social behaviour. I. Journal of Theoretical Biology 7, 1–16.
Hammerstein, P., 2003. Why is reciprocity so rare in social animals? a protestant appeal. In: Hammerstein, P. (Ed.), The Genetical and Cultural Evolution of Cooperation. MIT Press, Cambridge, pp. 83–93.
Hess, N.H., Hagen, E.H., 2006. Psychological adaptations for assessing gossip veracity. Human Nature 17 (3), 337–354.
Kandori, M., 1992. Social norms and community enforcement. The Review of Economic Studies 59 (1), 63–80.
Leimar, O., Hammerstein, P., 2001. Evolution of cooperation through indirect reciprocity. Proceedings of the Royal Society B 268, 745–753.
Mashima, R., Takahashi, N., 2008. The emergence of generalized exchange by indirect reciprocity. In: Biel, A., Eek, D., Garling, T., Gustafson, M. (Eds.), New

Issues and Paradigms in Research on Social Dilemmas. Springer, pp. 159–176 Chapter 10.
Maynard Smith, J., 1982. Evolution and the Theory of Games. Cambridge University Press, Cambridge.
Nakamaru, M., Kawata, M., 2004. Evolution of rumours that discriminate lying defectors. Evolutionary Ecology Research 6, 261–283.
Nowak, M.A., Sigmund, K., 1993. A strategy of win-stay lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. Nature 364, 56–58.
Nowak, M.A., Sigmund, K., 1998. The dynamics of indirect reciprocity. Journal of Theoretical Biology 194, 561–574.
Nowak, M.A., Sigmund, K., 2005. Evolution of indirect reciprocity. Nature 437, 1291–1298.
Ohtsuki, H., Iwasa, Y., 2004. How should we define goodness?—reputation dynamics in indirect reciprocity. Journal of Theoretical Biology 231, 107–120.
Pacheco, J.M., Santos, F.C., Chalub, F.A., 2006. Stern-judging: a simple, successful norm which promotes cooperation under indirect reciprocity. PLoS Computational Biology 2 (12), 1634–1638.
Panchanathan, K., Boyd, R., 2003. A tale of two defectors: the importance of standing for evolution of indirect reciprocity. Journal of Theoretical Biology 224 (1), 115–126.
Scheuring, I., 2009. Evolution of generous cooperative norms by cultural group selection. Journal of Theoretical Biology 257, 397–407.
Sommerfeld, R.D., Krambeck, H.-J., Milinski, M., 2008. Multiple gossip statements and their effect on reputation and trustworthiness. Proceedings of the Royal Society B 275, 2529–2536.
Takahashi, N., Mashima, R., 2006. The importance of subjectivity in perceptual errors on the emergence of indirect reciprocity. Journal of Theoretical Biology 243, 418–436.
Uchida, S., Sigmund, K., 2010. The competition of assessment rules for indirect reciprocity. Journal of Theoretical Biology 263, 13–19.