Involuntary defection and the evolutionary origins of empathy

Michael A. Fishman*

Department of Zoology, Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel

Received 30 November 2005; received in revised form 8 March 2006; accepted 6 May 2006
Available online 19 May 2006

Abstract

An occasional involuntary defection (IVD)—being unable, rather than unwilling, to donate help to others—is an intrinsic attribute of reciprocal cooperation (2003. J. Theor. Biol. 225–285). In fact, it is easy to see that—barring special circumstances—individuals that can donate help whenever requested do not need help of the same kind from others. That is, it is by no means clear why such individuals should participate in symmetric reciprocity interactions.

In this paper, I analyse the consequences of introducing IVD into direct reciprocity modeling and show that a simple form of empathy—not retaliating after being punished for IVD—is a prerequisite for evolutionarily stable cooperation. Furthermore, we will see that the stability of this, empathic retaliator, strategy increases with the number of opportunities for cooperative exchanges in the life of an average individual.

Keywords: Cooperation; Empathy; Evolutionary; ESS; Game Theory; Markov Processes; Reciprocity; Retaliation; Stability

0. Introduction

Humans are the ultimate social species. Thus, researching the evolution of sociality—besides its merits as an intellectual challenge—holds a promise of a better understanding of our own societies.

The investigation of reciprocal cooperation by game theoretical methods—reciprocity research, is motivated by the view that cooperation is an important component of sociality. In fact, it was argued that reciprocal cooperation is the evolutionary basis of ethics (cf. Alexander, 1987).

Modern reciprocity research was launched by Trivers (1971), who used iterated prisoner's dilemma to show that an individual may help an unrelated conspecific provided: (i) the cost to the donor is less than the benefit to the recipient and (ii) the favor is likely to be returned at a later date. That is, cooperation is enduring only if individuals receive more than they donate.

At present, symmetric reciprocity research (i.e. favors of the same kind are donated and solicited) falls into two main categories. Direct reciprocity (cf. Trivers, 1971; Axelrod and Hamilton, 1981; Brembs, 1996) deals with repeated exchanges of favors between a pair of individuals. Indirect reciprocity (cf. Alexander, 1979; Nowak and Sigmund, 1998a, b) deals with situations wherein cooperation is channeled toward the cooperative members of the community.

In both cases, the cooperation is conditional on the help-seeking individuals’ past behavior. That is, in direct reciprocity, conditional cooperators refuse to help those who denied them help previously, whereas in indirect reciprocity conditional cooperators refuse to help those whom they have observed to deny help to others. This is due to the fact that cooperators must protect themselves from exploitation by individuals who solicit favors but do not reciprocate, and the only possible guide to classifying a help-soliciting individual is its past behavior (Trivers, 1971).

Classical reciprocity research focused on the rules governing the assessment of the help-soliciting individuals by the conditional cooperators—decision rules, while assuming that a generic player is able to donate help whenever required by the pertinent decision rules. However, this approach neglects a crucial attribute of reciprocal cooperation.
Briefly: As discussed, (symmetric) reciprocal cooperation consists of helping others in order to secure their help (of the same kind) when needed. That is, these interactions are only beneficial to individuals who occasionally need help from others. However, when an individual needs help, it is unlikely to be able to donate help. Thus, an occasional involuntary defection (IVD)—being unable, rather than unwilling, to donate help—is an integral feature of symmetric reciprocal cooperation.

Although the idea of IVD was initially introduced in the context of indirect reciprocity (Fishman, 2003), as we shall see below, it has far-reaching consequences in the context of direct reciprocity.

The principal consequence of introducing IVD into direct reciprocity is that unconditional retaliation (i.e., every defection by an opponent triggers a retaliatory defection) a common feature of the classical conditional cooperation strategies (CCS) of direct reciprocity (cf. Brembs, 1996), becomes counterproductive. To wit: two players of an unconditionally retaliatory CCS cannot sustain cooperation past the first IVD, since it provokes a retaliatory defection, which brings on a retaliatory defection, etc. . . .

One way to avoid such cycles of retaliation is to formulate a CCS that does not retaliate every time its opponent defects. However, it is easy to see that such a “borebearing” retaliator is vulnerable to exploitation by cheaters. Similarly, a CCS that attempts to distinguish between a voluntary and involuntary defection and only punishes the former—a “sympathetic” retaliator; may stimulate the evolution of the ability to counterfeit incapacity and/or can be exploited by individuals who need help from others but are incapable of reciprocating (cf. Lotem et al., 1999).

However, while an individual cannot be certain of correctly classifying others’ behavior, it can be certain of correctly classifying its own. And, as we shall see below, not retaliating whenever the opponent’s defection was preceded by one’s own IVD, assures evolutionarily stable cooperation.

We shall call this CCS empathic retaliator (ER) because ER strategists use their own experience to make inferences about the behavior of others, and do not retaliate when an opponent acts as they would in similar circumstances (cf. Thompson, 1995).

Below I show that the ER strategy is evolutionarily stable under biologically relevant conditions. Moreover, we shall see that ER’s domains of stability, in both the parameter space and the strategies’ set, increase with the number of the occasions for cooperation in a generic lifespan.

This paper is organized as follows. In Section 1, I show that the ER strategy is evolutionarily stable in both the presence of the unconditional reciprocity strategies, and its own reduction mutants—mutation-wise robustness. In Section 2, we address the non-empathic approaches to reciprocity with IVD in detail. Finally, to facilitate the cogency of the presentation, the technical details of the derivation of the pertinent payoff matrixes are relegated to the appendix.

1. Properties of the ER strategy

We start by considering the competition between the ER and the unconditional reciprocity strategies. That is, we consider the interplay among three evolutionary strategies.

Unconditional Altruists grant requests for help whenever they are able.

Unconditional Defectors solicit help from others but never donate it.

Empathic Retaliators punish the opponent for every defection, unless it was preceded by their own IVD.1

Let us define the probability that an average individual is able to donate help upon request by \(0 < q < 1\), the relative cost of donating help by \(0 < c < 1\), and the benefits of receiving help by unity. That is, costs and benefits of an individual exchanges are scaled in terms of the benefits, which are assumed to exceed the costs (Trivers, 1971).

To calculate the payoffs for the game, I shall adopt the approach due to Nowak and Sigmund (1998b) wherein modeling of reciprocity interactions is simplified by assuming that the participants take turns soliciting and being solicited for help. Furthermore, I shall follow Brandt and Sigmund (2004) in assuming that the incidence of such cycles-of-interaction in a generic lifetime can be approximated by a Poisson distribution.

Remark. Although the cycles’ formulation may seem as an extreme simplification, it has a straightforward biological rationale. To wit, helping others in hope of a future reciprocation is intrinsically risky. Thus, a CCS that minimizes the risks to the unavoidable by refusing to help opponents who already received help and have not yet reciprocated, is less vulnerable to exploitation than more “patient” CCS. In other words, over the evolutionary time spans, the former will displace the later.

Reciprocity interactions in the presence of IVD are a Markov process with (i) countable number of steps—each step corresponding to one cycle of interactions; and (ii) a finite number of discrete states with a constant payoff associated with each state. Using the standard methods of analysing such processes (cf. Bharrucha-Reid, 1988)—see the appendix for details, we obtain the following payoff

\[
\begin{align*}
& R = q - c, \quad U = q, \quad N = q - c, \\
& \tilde{R} = n, \quad \tilde{U} = 0, \quad \tilde{N} = q - c.
\end{align*}
\]

1The ER strategy can be thought of as the classical tit-for-tat (TFT), (Axelrod and Hamilton, 1981) plus the empathic constraints on retaliation.
matrix for the UA/ER/UD game:

\[
\begin{pmatrix}
\mu q(1-c) & q(\rho - 2\mu c)/2 & -\mu c \\
q(2\mu - \rho c)/2 & q\rho(1-c)/2 & -q\theta c/2 \\
\mu q & q\theta/2 & 0
\end{pmatrix}
\]

(1a)

Here \( \mu \) is the parameter of the pertinent Poisson distribution:

\[
\theta = \frac{1 - e^{-\mu}}{\mu} \quad \text{and} \quad \rho = \frac{2\mu}{2 - q} + \frac{1 - q}{2 - q}. \quad \text{(1b)}
\]

Since we model reciprocity, we must assume \( \mu > 0 \) implying \( 2\mu > \max(\rho, \theta) \). Hence, UA strategy is strictly dominated by ER strategy, and can be excluded without changing the game (cf. Weibull, 1996).

Exclusion of UA, rearrangement, and multiplication by \( 2/q \), yield

\[
\begin{pmatrix}
\text{ER} & \text{UD} \\
\text{ER} & \begin{pmatrix} \rho(\gamma - c) & 0 \\
0 & \theta c \end{pmatrix} \\
\text{UD} & 0
\end{pmatrix}
\]

(2)

where \( \gamma = 1 - \theta/\rho \).

Let us define the frequency of the ER strategists in a population by \( 0 \leq x \leq 1 \). Then the replicator equation (cf. Weibull, 1996) for \( x(t) \) is given by

\[
x' = [\rho(\gamma - c) + \theta c](x - \xi)x(1-x),
\]

(3)

where \( \xi = \theta c/[\theta c + \rho(\gamma - c)] \).

We see that system (3) has two convergence points \( x = 0 \) and 1. Thus, system (2), and hence system (1), has two evolutionarily stable strategy (ESS) solutions.

- UD (\( x = 0 \)) is an ESS for all parameter values,
- ER (\( x = 1 \)) is an ESS, whenever \( c < \gamma \). (\( 0 < \xi < 1 \)).

Now, \( \gamma \) is monotone increasing in \( \mu \), and \( \gamma \to 1 \) as \( \mu \to \infty \). Thus, since \( 0 < c < 1 \), ER is an ESS whenever \( \mu \) is sufficiently large (Fig. 1).

Moreover, \( \xi \) is monotone decreasing in \( \mu \) and \( \xi \to 0 \) as \( \mu \to \infty \) (Fig. 2).

That is, when \( \mu \) is large; ER strategy is, effectively, the global attractor on the game’s strategy set. In other words, ER is the only likely ESS, if the expected number of the cycles of interaction per lifespan is sufficiently high.

By itself the fact that ER is an ESS of the UA/ER/UD game, is not enough to establish it as a biologically relevant CCS.

Briefly: Evolutionary strategies are specified by genes, which can be inactivated by mutations—yielding reduction mutants of that strategy. Thus, an evolutionary strategy can be considered adaptive only if it is mutation-wise robust i.e. is an ESS in the presence of its own reduction mutants (Selten and Hammerstein, 1984).

It is easy to see that in the specific case of ER, there are two possibilities for such reduction mutations.

- an ER may lose the ability to remember previous interactions—yielding a UA strategist,
- an ER may lose its “empathic” ability while retaining its ability to remember previous interactions—yielding a TFT player.

As shown above, UA mutants will be eliminated because they will have lower fitness than ER strategists. Thus, it remains to see under which conditions a population of ER strategists can be invaded by TFT mutants.

In the appendix, I show that the payoff for a TFT playing against an ER, \( \pi_{\text{UE}} \), is given by

\[
\pi_{\text{UE}} = \frac{1}{2} q \left[ 1 + \frac{1}{q} \omega_1 - \frac{(1 + q^2)\omega_1 + (1 - q)^2\omega_2}{1 - q + q^2} \right]
\]

where \( \omega_1 = \frac{1 - e^{-q(1-q)\mu}}{1 - q} \) and \( \omega_2 = \frac{1 - e^{-\mu}}{1 - q} \). \quad \text{(4a)}

Now, the payoff for ER vs. ER interactions is \( \pi_{22} = q\rho(1-c)/2 \). Hence, TFT mutants cannot invade a
population of ER strategists if \( \pi_{\text{UE}} < \pi_{22} \). But, \( \pi_{22} - \pi_{\text{UE}} = (q/2)\phi(\gamma' - c) \) where

\[
\phi = \rho - (1 + q^2)\omega_1 + (1 - q)^2\omega_2 > 0
\]

and

\[
\gamma' = 1 - \frac{(1 - q)[\omega_1 - q(1 - q)\omega_2]}{q(1 - q + q^2)\phi}.
\]  

(4b)

Since \( 0 < \gamma' < \gamma \) for any finite \( \mu \), the probability that ER is an ESS of the UA/ER/UD game but is not of the UA/ER/TFT/UD game, is given by

\[
\delta(\mu) = 1 - \frac{\gamma'}{\gamma}.
\]  

(4c)

However, \( \gamma' \) is monotone increasing in \( \mu \), and \( \gamma' \to 1 \) as \( \mu \to \infty \). Thus, \( \delta \) decreases as \( \mu \) increases (Fig. 3).

**Remark.** The fact that ER can mutate into TFT does not necessarily mean that TFT is an evolutionary precursor of ER. In fact it is easy to show, see below, that UD is the only ESS of the UA/TFT/UD game with IVD, except under an unrealistically restrictive set of conditions. Thus, other considerations aside, TFT is unlikely to have had persisted long enough to give rise to ER.

2. **Non-empathic approaches to reciprocity with IVD**

In the introduction, I mention the drawbacks of unconditional retaliation in IVD context, and briefly discuss two non-empathic approaches to overcoming these drawbacks.

- **forbearing retaliator** approach i.e. a CCS that does not retaliate every time its opponent defects,
- **sympathetic retaliator** approach i.e. a CCS that attempts to distinguish between a voluntary and involuntary defection and only punishes the former.

Below, these possibilities are addressed in more detail. I start with analysing the UA/TFT/UD game in the presence of IVD. This analysis serves two purposes: (i) the TFT strategy serves to represents the drawbacks of unconditional retaliation in IVD context, and (ii) a spectrum of strategies mixing UA and TFT serves to represent the possibility of the forgiving retaliators, without neglecting unconditional cooperation.\(^3\)

\(^2\)To recollect (cf. Wikipedia). *Empathy* is the ability to ‘put oneself into another’s shoes’. Whereas, *sympathy* is the feeling of compassion for others i.e., “feeling sorry” for them.

\(^3\)As discussed in Section 1, the presence of UA is a prerequisite for biological realism in reciprocity modeling.
$\mu$ is large. In other words, we can assume that neither TFT, nor its “forbearing” alternatives are ESS in the presence of IVD and reduction mutations.

To address the issue of the sympathetic retaliation (SR), we consider a game between SR strategists and an evolutionary strategy the players of which solicit help but are incapable of reciprocating—let us denote this strategy by GI for genuine incapacity.\footnote{That is, GI strategy is a heritable analog of the \textit{phenotypic defectors} (Lotem et al., 1999).}

For an example of such a strategy, consider species/population where reciprocal cooperation consists of food sharing, and assume some individuals bear genes that make them incapable of foraging effectively.\footnote{Obviously, such a situation does not merit a consideration—except in the presence of SR.}

It is easy to see that the payoff matrix for the SR/GI game (where, for the sake of simplicity, SR strategists are assumed to have perfect discrimination between a voluntary and involuntary defection) is identical to the payoff matrix for the UA/UD game—see Eq. (5).

### 3. Concluding remarks

After Selten and Hammerstein (1984) have demonstrated that TFT is not an ESS under biologically relevant conditions, a considerable amount of attention was focused on the effects of noise—\textit{errors in implementation}, on the decision rules governing CCS responses (cf. Boyd, 1989; Boerlijst et al., 1997). This paper cannot be complete without a brief discussion of the differences between the “noise” approach and the one presented in this paper.

\textit{Briefly:} There are two main differences.

- Errors in implementation are, necessarily, bi-directional i.e. CCS formulated to address this phenomenon help “underserving”, as well as denying help to the “deserving”. This, inevitably, leads to decision rules that are substantially more complex than that of the ER strategy. In particular, such CCS are subject to error in interpreting their opponents’ moves—\textit{perception}. And an error in perception terminates cooperation (cf. Boerlijst et al., 1997).

- More importantly, noise in CCS players’ responses \textit{does not} affect unconditional cooperators. However, genes responsible for conditionality in conditional cooperation, are subject to mutational inactivation—and therefore, unconditional cooperators are an integral component of biologically relevant reciprocity models (Selten and Hammerstein, 1984). Thus, reciprocity models \textit{sans} IVD are either biased in favor of unconditional cooperators, or are biased by omitting unconditional cooperators.

The preceding discussion should not be taken to mean that the examination of the effect of noise in the IVD/empathic retaliation context is not a viable direction for a future research. For now, however, I wish to summarize the current results.

I have argued that IVD is an integral component of reciprocal cooperation (Tooby and Cosmides, 1996; Fishman, 2003). In fact, it is easy to see that, barring special circumstances, individuals that can always donate help do not need help (of the same kind) from others. That is, within the constraints of symmetric reciprocity models, the optimum state for such individuals is universal defection. Consequently, IVD is a crucial component of reciprocity modeling.

Once we acknowledge the existence of IVD, we see that a simple form of empathy is a prerequisite for enduring cooperation in direct reciprocity context. In retrospect, this should not be surprising, as it was suggested (cf. Page and Nowak, 2002; Preston and de Waal, 2002) that cooperation, and sociality in general, cannot function without empathy. Consequently, the very simplicity of the empathic component of the ER strategy marks it as a possible evolutionary precursor of the observable empathic components of sociality.

### Acknowledgment

This paper is dedicated to my late mentor Lee Segel, who taught me that “reality-check” is the most important tool of a mathematical modeler.

### Appendix

Reciprocity interactions between a pair of players are \textit{initiated} when one individual solicits help from another for the first time. In the presence of IVD, these interactions are stochastic; yielding \textit{Markov chains} with countable number of steps—each step corresponding to one round of interactions—and finite number of possible discrete states
where the payoff associated with each discrete state is constant.

The most important attribute of such processes is that the probability of transition from any state to any other state does not change with time (cf. Bharucha-Reid, 1988). Thus, let us define:

(a) The initial probability distribution on states = 
\[ (p_1(1), \ldots, p_m(m))^T. \]
(b) The state specific payoffs = \( (w_1, \ldots, w_m))^T, \)
(c) The transition probabilities matrix (TPM), \( T = [p_{ij}], \)

where \( p_{ij}; i,j \in \{1, \ldots, m\}; \) is the probability of the transition from state \( j \) to state \( i. \)

Remark. When \( m = 2, \) we can use the fact that \( p_n(2) = 1 - p_n(1), \) and therefore \( p_n(1) = (p_{11} - p_{12})p_{n-1}(1) \) + \( p_{12}, \) to derive \( p_n(1) \) by recursion. However, given access to a computer–algebra software package (e.g. Maple), it is easier to calculate \( T^{n-1}. \)

Since there are size considerations, I shall not detail all the calculation pertinent to the current paper. Rather, I shall illustrate the methodology by detailing the most complex case: ER vs. TFT interactions.

As discussed, we define the probability of IVD by \( 1 - q, \) and scale costs and benefits in terms of benefits i.e. the relative costs of donating help are 0 < \( c < 1, \) and the benefits of receiving help are unity.

There are two possibilities: ER initiates the interactions, or TFT initiates the interactions. Let ER responses be dashed, TFT responses be solid, and let RD stand for retaliatory defection.

(I) If ER initiates the interactions, then there are three possible states (Fig. 5).

![Fig. 5. A schematic representation of the TFT vs. ER interactions initiated by ER.](image)

Hence, the TPM is given by

\[
T = \begin{pmatrix}
q^2 & q & 0 \\
q(1-q) & 1-q & 0 \\
1-q & 0 & 1 \\
\end{pmatrix}.
\] (A.1a)

And therefore, since \( p_1(1) = 1: \)

\[
\begin{pmatrix}
p_{k}(1) \\
p_{k}(2) \\
p_{k}(3) \\
\end{pmatrix} = [T^{k-1}]_{11} = \begin{pmatrix}
q^2(1-q + q^2)^{k-2} \\
q(1-q)(1-q + q^2)^{k-2} \\
1-q(1-q + q^2)^{k-2} \\
\end{pmatrix}. \] (A.1b)

Note: \( \partial p_k(3)/\partial k > 0 \) and \( p_k(3) \rightarrow 1. \) That is, TFT vs. ER cooperation, initiated by ER, is not sustainable.

To derive the state- and strategy-specific payoffs, we use the same scheme, but replace the next state's designation with the strategy-specific payoffs (i.e. a “\( q, \)” move by the opponent = 1, a “\( q, \)” move by the focal = \( -c, \) and the “\( 1-q, \)” moves = 0).

In these terms, the state-specific TFT payoffs are given by

\[
u_{TFT}(1) = q(1-q), \quad u_{TFT}(2) = q, \quad u_{TFT}(3) = 0. \] (A.1c)

Thus, the strategy-specific payoffs for \( n \) cycles of interaction are given by

\[
u_{TFT}^n = q \frac{1 - (1-q + q^2)^n}{1-q} - q \frac{1-q(1-q + q^2)^{n-1}}{1-q} \cdot \] (A.1d)

(II) If TFT initiates the interactions, then there are two possible states (Fig. 6).

Hence, the TPM is given by

\[
T = \begin{pmatrix}
1-q+q^2 & 0 \\
1-q & 1 \\
\end{pmatrix}. \] (A.2a)

And therefore,

\[
p_1(1) = (1-q + q^2)^{k-1}. \] (A.2b)
Note: $p_k(2) = 1 - p_k(1)$. Hence, $\frac{\partial p_k(2)}{\partial k} > 0$ and $p_k(2) \rightarrow 1$. That is, TFT vs. ER cooperation, initiated by TFT, is not sustainable.

Proceeding as in Part I, we see that the state-specific TFT payoffs are

$$v_{TFT}(1) = q(1 - qc), \quad v_{TFT}(2) = 0.$$ (A.2c)

Thus, the strategy-specific payoffs for $n$ cycles of interaction are

$$v_n^{TFT} = \frac{1 - (1 - q + q^2)^n}{1 - q} (1 - qc).$$ (A.2d)

Hence, the per life-span payoff for a TFT playing against an ER is given by

$$\pi_{UE} = \sum_{n=1}^{\infty} \frac{1}{n!} \left[ v_n^{TFT} + v_n^{ER} \right] \frac{\mu^n}{n!} e^{-\mu}$$

$$= \frac{1}{2} q \left[ 1 + q - \omega_1 - \frac{(1 + q^2)\omega_1 + (1 - q)^2\omega_2}{1 - q + q^2} c \right],$$ (A.3a)

where

$$\omega_1 = \frac{1 - e^{-(1-q)\mu}}{1 - q} \quad \text{and} \quad \omega_2 = \frac{1 - e^{-\mu}}{1 - q}.$$