



Direct reciprocity with costly punishment: Generous tit-for-tat prevails

David G. Rand^{a,b,*}, Hisashi Ohtsuki^{a,c}, Martin A. Nowak^{a,d,e}

^a Program for Evolutionary Dynamics, Harvard University, Cambridge MA 02138, USA

^b Department of Systems Biology, Harvard University, Cambridge, MA 02138, USA

^c Tokyo Institute of Technology, Muguro-ku, Tokyo 152-8552, Japan

^d Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA

^e Department of Mathematics, Harvard University, Cambridge, MA 02138, USA

ARTICLE INFO

Article history:

Received 3 March 2008

Received in revised form

27 August 2008

Accepted 18 September 2008

Available online 2 October 2008

Keywords:

Cooperation

Punishment

Evolution

Nash equilibrium analysis

Finite population size analysis

Computer simulation

Reciprocity

ABSTRACT

The standard model for direct reciprocity is the repeated Prisoner's Dilemma, where in each round players choose between cooperation and defection. Here we extend the standard framework to include costly punishment. Now players have a choice between cooperation, defection and costly punishment. We study the set of all reactive strategies, where the behavior depends on what the other player has done in the previous round. We find all cooperative strategies that are Nash equilibria. If the cost of cooperation is greater than the cost of punishment, then the only cooperative Nash equilibrium is generous-tit-for-tat (GTFT), which does not use costly punishment. If the cost of cooperation is less than the cost of punishment, then there are infinitely many cooperative Nash equilibria and the response to defection can include costly punishment. We also perform computer simulations of evolutionary dynamics in populations of finite size. These simulations show that in the context of direct reciprocity, (i) natural selection prefers generous tit-for-tat over strategies that use costly punishment, and (ii) that costly punishment does not promote the evolution of cooperation. We find quantitative agreement between our simulation results and data from experimental observations.

Published by Elsevier Ltd.

1. Introduction

Two key mechanisms for the evolution of any cooperative (or 'pro-social' or 'other-regarding') behavior in humans are direct and indirect reciprocity. Direct reciprocity means there are repeated encounters between the same two individuals, and my behavior towards you depends on what you have done to me. Indirect reciprocity means there are repeated encounters in a group of individuals, and my behavior towards you also depends on what you have done to others. Our social instincts are shaped by situations of direct and indirect reciprocity. All of our interactions have possible consequences for the future. Either I might meet the same person again or others might find out what I have done and adjust their behavior towards me. Direct reciprocity has been studied by many authors (Trivers, 1971; Axelrod and Hamilton, 1981; Axelrod, 1984; Selten and Hammerstein, 1984; Nowak and Sigmund, 1989, 1992, 1993; Kraines and Kraines, 1989; Fudenberg and Maskin, 1990; Imhof et al., 2005, 2007). For indirect reciprocity, see (Sugden, 1986; Alexander, 1987; Kandori, 1992; Nowak and Sigmund, 1998, 2005; Lotem

et al., 1999; Ohtsuki and Iwasa, 2004, 2005; Panchanathan and Boyd, 2005; Brandt and Sigmund, 2006; Pacheco et al., 2006a).

Much of human history was spent in small groups, where people knew each other. In such a setting direct and indirect reciprocity must occur. Therefore, even if we think of group selection as a mechanism for the evolution of cooperation among humans (Wynne-Edwards, 1962; Wilson, 1975; Boyd and Richerson, 1990; Wilson and Sober, 1994; Bowles, 2001; Nowak, 2006; Traulsen and Nowak, 2006) this could only occur in combination with direct and indirect reciprocity. Reciprocity is an unavoidable consequence of small group size, given the cognitive abilities of humans.

Yamagishi (1986, 1988) introduced experimental studies of costly punishment among humans. Axelrod (1986) suggested that costly punishment can stabilize social norms. Fehr and Gächter (2000, 2002) suggested that costly punishment is an alternative mechanism for the evolution of human cooperation that can work independently of direct or indirect reciprocity. While it is interesting to study the effect of costly punishment on human (or animal) behavior (Ostrom et al., 1994; Clutton-Brock and Parker, 1995; Burnham and Johnson, 2005; Güerker et al., 2006; Rockenbach and Milinski, 2006; Dreber et al., 2008; Herrmann et al., 2008; Sigmund, 2008), it is not possible to consider costly punishment as an independent mechanism. If I punish you because you have defected with me, then I use direct reciprocity.

* Corresponding author at: Program for Evolutionary Dynamics, One Brattle Square, Harvard University, Cambridge, MA 02138, USA.

E-mail address: drand@fas.harvard.edu (D.G. Rand).

If I punish you because you have defected with others, then indirect reciprocity is at work. Therefore most models of costly punishment that have been studied so far (Boyd and Richerson, 1992; Sigmund et al., 2001; Boyd et al., 2003; Brandt et al., 2003; Fowler, 2005; Nakamaru and Iwasa, 2006; Hauert et al., 2007) tacitly use direct or indirect reciprocity.

Costly punishment is sometimes called ‘altruistic punishment’, because some people use it in the second and last round of a game where they cannot directly benefit from this action in the context of the experiment (Fehr and Gächter, 2002; Boyd et al., 2003). We find the term ‘altruistic punishment’ misleading, because typically the motives of the punishers are not ‘altruistic’ and the strategic instincts of people are mostly formed by situations of repeated games, where they could benefit from their action. It is more likely that punishers are motivated by internal anger (Kahneman et al., 1998; Carlsmith et al., 2002; Sanfey et al., 2003) rather than by the noble incentive to do what is best for the community. Thus, ‘costly punishment’ is a more precise term than ‘altruistic punishment’. Costly punishment makes no assumptions about the motive behind the action.

Since costly punishment is a form of direct or indirect reciprocity, the suggestion that costly punishment might promote human cooperation must be studied in the framework of direct or indirect reciprocity. Here we attempt to do this for direct reciprocity.

One form of direct reciprocity is described by the repeated Prisoner’s Dilemma. In each round of the game, two players can choose between cooperation, *C*, and defection, *D*. The payoff matrix is given by

$$\begin{matrix} & C & D \\ C & (a_2 & a_4) \\ D & (a_1 & a_3) \end{matrix} \quad (1)$$

The game is a Prisoner’s Dilemma if $a_1 > a_2 > a_3 > a_4$.

We can also say that cooperation means paying a cost, *c*, for the other person to receive a benefit *b*. Defection means either ‘doing nothing’ or gaining payoff *d* at the cost *e* for the other person. In this formulation, the payoff matrix is given by

$$\begin{matrix} & C & D \\ C & (b - c & -c - e) \\ D & (d + b & d - e) \end{matrix} \quad (2)$$

We have $b > c > 0$ and $d, e \geq 0$. This payoff matrix is a subset of all possible Prisoner’s Dilemmas. Not every Prisoner’s Dilemma can be written in this form, only those that have the property of ‘equal gains from switching’, $a_1 - a_2 = a_3 - a_4$ (Nowak and Sigmund, 1990).

Including costly punishment means that we have to consider a third strategy, *P*, which has a cost α for the actor and a cost β for the recipient. The 3×3 payoff matrix is of the form

$$\begin{matrix} & C & D & P \\ C & (b - c & -c - e & -c - \beta) \\ D & (d + b & d - e & d - \beta) \\ P & (-\alpha + b & -\alpha - e & -\alpha - \beta) \end{matrix} \quad (3)$$

Note that the idea of ‘punishment’ is not new in the world of the Prisoner’s Dilemma. The classical ‘punishment’ for defection is defection. Tit-for-tat punishes defection with defection. The new proposal, however, is that there is another form of punishment which is costly for the punisher. Two questions then present themselves. Is it advantageous to use costly punishment, *P*, instead of defection, *D*, in response to a co-player’s defection? And furthermore, does costly punishment allow cooperation to succeed in situations where tit-for-tat does not? We will explore these questions in the present paper.

Section 2 contains an analysis of Nash equilibria among reactive strategies. Section 3 presents the results of computer simulations. Section 4 compares our theoretical findings with experimental data. Section 5 concludes.

2. Nash-equilibrium analysis

We are interested in Nash equilibria of the repeated game given by the payoff matrix (3). We assume $b, c, \alpha, \beta > 0$ and $d, e \geq 0$ throughout the paper. We refer to one game interaction as a ‘round’. With probability w ($0 < w < 1$), the game continues for another round. With probability $1 - w$, the game terminates. The number of rounds follows a geometrical distribution with mean $1/(1 - w)$. The parameter w can also be interpreted as discounting future payoffs.

A ‘strategy’ of a player is a behavioral rule that prescribes an action in each round. We assume that each player has a probabilistic strategy as follows. In the first round, a player chooses an action (either *C*, *D*, or *P*) with probability p_0, q_0 and r_0 , respectively. From the second round on, a player chooses an action depending on the opponent’s action in the previous round. The probability that a player chooses *C*, *D*, or *P*, is given by p_i, q_i , and r_i , for each possible previous action ($i = 1, 2, 3$ for *C, D, P*) of the opponent. Thus a strategy is described by 12 values as

$$s = \begin{matrix} & C & D & P \\ \text{Initial move} & (p_0 & q_0 & r_0) \\ \text{Response to C} & (p_1 & q_1 & r_1) \\ \text{Response to D} & (p_2 & q_2 & r_2) \\ \text{Response to P} & (p_3 & q_3 & r_3) \end{matrix} \quad (4)$$

Since p_i, q_i, r_i are probabilities, our strategy space is

$$S_3^4 = \prod_{i=0}^3 \{(p_i, q_i, r_i) \mid p_i + q_i + r_i = 1, p_i, q_i, r_i \geq 0\}. \quad (5)$$

This is the product of four simplexes, S_3 . Note that we are considering the set of reactive strategies (Nowak, 1990): a player’s move only depends on the co-player’s move in the previous round. This strategy space includes not only reciprocal strategies, but also non-reciprocal unconditional strategies ($p_i = p, q_i = q, r_i = r$), and paradoxical ones that cooperate less with cooperators than with defectors or punishers (Herrmann et al., 2008). For example, the strategy ‘always defect’ (ALLD) is given by $p_i = 0, q_i = 1, r_i = 0$. Its action does not depend on the opponent’s behavior.

We introduce errors in execution. It is well-known that errors play an important role in the analysis of repeated games (Molander, 1985; May, 1987; Nowak and Sigmund, 1989, 1992, 1993; Fudenberg and Maskin, 1990; Fudenberg and Tirole, 1991; Lindgren, 1991; Lindgren and Nordahl, 1994; Boerlijst et al., 1997; Wahl and Nowak, 1999a, b). In our model, a player fails to execute his intended action with probability 2ε . When this occurs, he does one of the other two unintended actions randomly, each with probability ε . We assume $0 < \varepsilon < \frac{1}{3}$.

We will now calculate all Nash equilibria of the repeated game. Let $u(s_1, s_2)$ represent the expected total payoff of an s_1 -strategist against an s_2 -strategist. Strategy s is a Nash equilibrium of the repeated game if the following inequality holds for any $s' \in S_3^4$:

$$u(s, s) \geq u(s', s). \quad (6)$$

This condition implies that no strategy s' can do better than strategy s against s . In Appendix A, we show a complete list of Nash equilibria.

Since we are interested in the evolution of cooperation, we will restrict our attention to ‘cooperative’ Nash equilibria. We define

a Nash-equilibrium strategy, s , as *cooperative* if and only if two s -strategists always cooperate in the absence of errors (i.e. $\varepsilon \rightarrow 0$).

It is easy to show that the criterion above means that we search for Nash equilibria of the form

$$s = \begin{matrix} \text{Initial move} \\ \text{Response to C} \\ \text{Response to D} \\ \text{Response to P} \end{matrix} \begin{pmatrix} C & D & P \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ p_2 & q_2 & r_2 \\ p_3 & q_3 & r_3 \end{pmatrix}. \tag{7}$$

According to the results in Appendix A, there are three types of cooperative Nash equilibria as follows.

2.1. Cooperative Nash equilibria

2.1.1. Cooperative Nash equilibria without defection

Let $w' = w(1 - 3\varepsilon)$. If

$$w' > \frac{c+d}{b+\beta} \text{ and } \alpha \leq c \tag{8}$$

then there exist cooperative Nash equilibria where defection is never used. These strategies are of the form

$$s = \begin{matrix} \text{Initial move} \\ \text{Response to C} \\ \text{Response to D} \\ \text{Response to P} \end{matrix} \begin{pmatrix} C & D & P \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1-r_2 & 0 & r_2 \\ 1-r_3 & 0 & r_3 \end{pmatrix}. \tag{9}$$

The probabilities r_2 and r_3 must satisfy

$$r_2 > \frac{c+d}{w'(b+\beta)} \tag{10}$$

and

$$r_3 = \frac{c-\alpha}{w'(b+\beta)}. \tag{11}$$

2.1.2. Cooperative Nash equilibria without punishment

If

$$w' \geq \frac{c+d}{b+e} \tag{12}$$

then there exist cooperative Nash equilibria where punishment is never used. These strategies are of the form

$$s = \begin{matrix} \text{Initial move} \\ \text{Response to C} \\ \text{Response to D} \\ \text{Response to P} \end{matrix} \begin{pmatrix} C & D & P \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1-q_2 & q_2 & 0 \\ 1-q_3 & q_3 & 0 \end{pmatrix}. \tag{13}$$

The probabilities q_2 and q_3 must satisfy

$$q_2 = \frac{c+d}{w'(b+e)} \tag{14}$$

and

$$q_3 > \frac{c-\alpha}{w'(b+e)}. \tag{15}$$

Note that while q_2 must be a specific value, q_3 must only be greater than a certain threshold.

2.1.3. Mixed cooperative Nash equilibria

If

$$w' \geq \frac{c+d}{b+\max\{\beta, e\}} \text{ and } \alpha \leq c \tag{16}$$

then there exist cooperative Nash equilibria where a mixture of defection and punishment can be used. These strategies are of the form

$$s = \begin{matrix} \text{Initial move} \\ \text{Response to C} \\ \text{Response to D} \\ \text{Response to P} \end{matrix} \begin{pmatrix} C & D & P \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ p_2 & q_2 & r_2 \\ p_3 & q_3 & r_3 \end{pmatrix}. \tag{17}$$

The probabilities $p_i, q_i,$ and r_i ($i = 2, 3$) must satisfy

$$bp_2 - eq_2 - \beta r_2 = b - \frac{c+d}{w'} \tag{18}$$

and

$$bp_3 - eq_3 - \beta r_3 = b - \frac{c-\alpha}{w'}. \tag{19}$$

2.1.4. Does punishment promote cooperation?

We can now ask if costly punishment allows cooperation to succeed when classical direct reciprocity alone does not. From Eqs. (8), (12), (16), we see that if the conditions

$$\frac{c+d}{b+\beta} \leq w' < \frac{c+d}{b+e} \tag{20}$$

and

$$\alpha \leq c \tag{21}$$

hold, there exist cooperative Nash equilibria that use punishment, but none that use only cooperation and defection. Therefore costly punishment can create cooperative Nash equilibria in parameter regions where there would have been none with only tit-for-tat style strategies.

In a classical setting where $d = e = 0$ and $\varepsilon \rightarrow 0$, there exist Nash equilibria that use only cooperation and defection if $w \geq c/b$. However, the condition is loosened to $w \geq c/(b + \beta)$ by introducing costly punishment if $\alpha \leq c$. In this case, costly punishment can allow cooperative Nash equilibria even when cooperation is not beneficial ($b < c$). When $\alpha > c$, on the other hand, there are no such cooperative Nash equilibria and punishment does not promote cooperation.

2.2. Equilibrium selection and the best cooperative Nash equilibrium

We will now characterize the strategy that has the highest payoff against itself and is a cooperative Nash equilibrium.

2.2.1. Punishment is cheaper than cooperation: $\alpha \leq c$

For $\alpha \leq c$, there is at least one cooperative Nash equilibrium if

$$w' \geq \frac{c+d}{b+\max\{\beta, e\}}. \tag{22}$$

We now ask at which cooperative Nash equilibrium is the payoff, $u(s, s)$, maximized, given that Eq. (25) is satisfied. Since each cooperative Nash equilibrium achieves mutual cooperation in the absence of errors, the payoff $u(s, s)$ is always of the form

$$u(s, s) = \frac{b-c}{1-w} - \mathcal{O}(\varepsilon). \tag{23}$$

Here $\mathcal{O}(\varepsilon)$ represents a term of order of ε or higher. We now compare the magnitude of this error term among our cooperative Nash equilibria.

Our calculation shows that all strategies given by Eqs. (17)–(19) are co-maximizers of the payoff. The maximum payoff is given by

$$u(s, s) = \frac{b - c}{1 - w} - \varepsilon \cdot \frac{2b + e + \beta}{1 - w}. \tag{24}$$

Figs. 1A and B are graphical representations of these ‘best’ cooperative Nash equilibria. Therefore, if $\alpha \leq c$ holds, then both defection and punishment can be used as a response to non-cooperative behavior.

2.2.2. Punishment is more expensive than cooperation: $\alpha > c$

If $\alpha > c$, then punishment-free strategies are the only candidates for cooperative Nash equilibria. The condition under which there is at least one cooperative Nash equilibrium is

$$w' \geq \frac{c + d}{b + e}. \tag{25}$$

When Eq. (25) is satisfied, we obtain cooperative Nash equilibria in the form of Eqs. (12) and (13). Calculation shows

that the payoff, $u(s, s)$, is maximized for

$$s = \begin{matrix} & C & D & P \\ \text{Initial move} & 1 & 0 & 0 \\ \text{Response to C} & 1 & 0 & 0 \\ \text{Response to D} & 1 - q_2 & q_2 & 0 \\ \text{Response to P} & 1 & 0 & 0 \end{matrix}, \tag{26}$$

where

$$q_2 = \frac{c + d}{w(b + e)}. \tag{27}$$

It is noteworthy that the strategy (26) corresponds to ‘generous tit-for-tat’ (Nowak, 1990; Nowak and Sigmund, 1992). Fig. 1C shows this ‘best’ cooperative Nash equilibrium. The maximum payoff is given by

$$u(s, s) = \frac{b - c}{1 - w} - \varepsilon \cdot \frac{2b - c + e + \alpha + \beta}{1 - w}. \tag{28}$$

Therefore, if $\alpha > c$ holds, then the ‘best’ strategy is to defect against a defector with probability (27) and to always cooperate otherwise. Using punishment either reduces the payoff or destabilizes this strategy.

2.2.3. Summary of equilibrium selection

Tables 1 and 2 summarize our search for the best cooperative Nash equilibria.

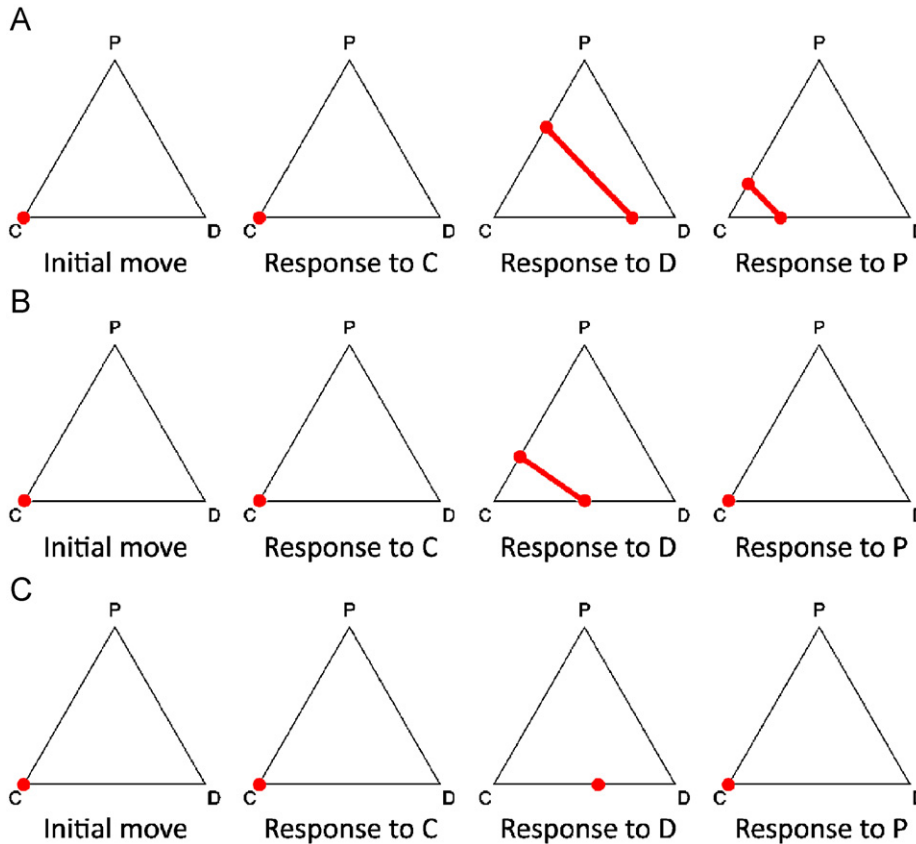


Fig. 1. The best (highest payoff) cooperative Nash equilibria can use costly punishment only if punishment is not more expensive than cooperation, $\alpha \leq c$. The strategy space of our model, S_3^2 is shown. From left to right, each simplex represents the initial action, reaction to cooperation, reaction to defection and reaction to punishment, respectively. Each corner (labeled C, D, P) represents a pure reaction. A point in the simplex represents a probabilistic reaction. (A) The best cooperative Nash equilibria when punishment is cheaper than cooperation can use punishment in response to defection and/or punishment. Any pair of points from the line in the D-simplex and the line in the P-simplex is a payoff maximizing cooperative Nash equilibrium. (B) The best cooperative Nash equilibria when punishment is equal in cost to cooperation can use punishment in response to defection, but always cooperates in response to punishment. (C) The best cooperative Nash equilibrium when punishment is more expensive than cooperation is generous tit-for-tat. Only defection and cooperation are used in reaction to defection. Punishment is never used.

Table 1
The ‘best’ cooperative Nash equilibria derived in Section 2.2

	Punishment is less costly than cooperation, $\alpha \leq c$	Punishment is more costly than cooperation, $\alpha > c$
Initial move	C	C
Response to C	C	C
Response to D	C or D or P Any (p_2, q_2, r_2) that satisfies $q_2 + \frac{b + \beta}{b + e} r_2 = \frac{c + d}{w(1 - 3\varepsilon)(b + e)}$	C or D $p_2 = 1 - \frac{c + d}{w(1 - 3\varepsilon)(b + e)}$, $q_2 = \frac{c + d}{w(1 - 3\varepsilon)(b + e)}$, $r_2 = 0$
Response to P	C or D or P Any (p_3, q_3, r_3) that satisfies $q_3 + \frac{b + \beta}{b + e} r_3 = \frac{c - \alpha}{w(1 - 3\varepsilon)(b + e)}$	C

Such strategies are optimal in that each receives the highest payoff against itself. If $\alpha \leq c$, optimal cooperative Nash equilibria exist which use costly punishment in response to defection and/or punishment. If $\alpha > c$, the unique optimal cooperative Nash equilibrium is generous tit-for-tat, which never uses costly punishment. It is interesting that as punishment becomes more expensive, the optimal response to an opponent’s punishment becomes more generous. If $\alpha \geq c$, punishment use is always responded to with full cooperation.

Table 2
The best (highest payoff) cooperative Nash equilibria derived in Section 2.2 when $\varepsilon \rightarrow 0$ and $d = e = 0$

	Punishment is less costly than cooperation, $\alpha \leq c$	Punishment is more costly than cooperation, $\alpha > c$
Initial move	C	C
Response to C	C	C
Response to D	C or D or P Any (p_2, q_2, r_2) that satisfies $q_2 + \frac{b + \beta}{b} r_2 = \frac{c}{bw}$	C or D $p_2 = 1 - \frac{c}{bw}$, $q_2 = \frac{c}{bw}$, $r_2 = 0$
Response to P	C or D or P Any (p_3, q_3, r_3) that satisfies $q_3 + \frac{b + \beta}{b} r_3 = \frac{c - \alpha}{bw}$	C

Again, costly punishment is only used if $\alpha \leq c$, and otherwise the optimal strategy is generous tit-for-tat. You can also see that if $\alpha \leq c$, strategies which respond to D using only C and P ($q_2 = 0$) are more forgiving than strategies which respond to D using only C and D ($r_2 = 0$).

3. Individual based simulations

Our analysis in the previous section indicates that there are infinitely many Nash equilibria in this system, some of which are cooperative, and some of which may or may not use costly punishment. We would like to know which strategies are selected by evolutionary dynamics in finite populations. In order to do this, we turn to computer simulations.

3.1. Simulation methods

We consider a well-mixed population of fixed size, N . Again we consider the set of all ‘reactive strategies’. For mathematical simplicity, we begin by assuming that games are infinitely repeated, $w = 1$. We later investigate the case $w < 1$. We only examine stochastic strategies, where $0 < p_i, q_i, r_i < 1$ for all i . Therefore, it is not necessary to specify the probabilities p_0, q_0 , and r_0 for the initial move if $w = 1$. For $w < 1$, we will assume that a strategy’s initial move is the same as its response to cooperation, $p_0 = p_1, q_0 = q_1, r_0 = r_1$.

A game between two players s_1 and s_2 can be described by a Markov process. For $w = 1$, the average payoff per round, $u(s_1, s_2)$, is calculated from the stationary distribution of actions (Nowak and Sigmund, 1990). For $w < 1$, the total payoff is approximated by truncating the series after the first 50 terms.

In our simulations, each player s_i plays a repeated Prisoner’s Dilemma with punishment against all other players. The average payoff of player s_i is given by

$$\pi_i = \frac{1}{N - 1} \sum_{\substack{j=1 \\ j \neq i}}^N u(s_i, s_j). \tag{29}$$

We randomly sample two distinct players $s^{(T)}$ (Teacher) and $s^{(L)}$ (Learner) from the population, and calculate the average payoffs for each. The learner then switches to the teacher’s strategy with probability

$$p = \frac{1}{1 + e^{-\frac{\pi^{(T)} - \pi^{(L)}}{\tau}}}. \tag{30}$$

This is a monotonically increasing function of the payoff-difference, $\pi^{(T)} - \pi^{(L)}$, taking the values from 0 to 1. This update rule is called the ‘pairwise comparison’ (Pacheco et al., 2006b; Traulsen et al., 2006, 2007). The parameter $\tau > 0$ in (30) is called the ‘temperature of selection’. It is a measure of the intensity of selection. For very large τ we have weak selection (Nowak et al., 2004).

In learning, we introduce a chance of ‘mutation’ (or ‘exploration’). When the learner switches his strategy, then with probability μ he adopts a completely new strategy. In this case, the probabilities p_i, q_i , and r_i for each i are randomly generated using a U-shaped probability density distribution as in Nowak and

Sigmund (1992, 1993). We use this distribution to increase the chance of generating mutant strategies that are close to the boundary of the strategy space. This makes it easier to overcome ALLD populations. See Appendix B for details.

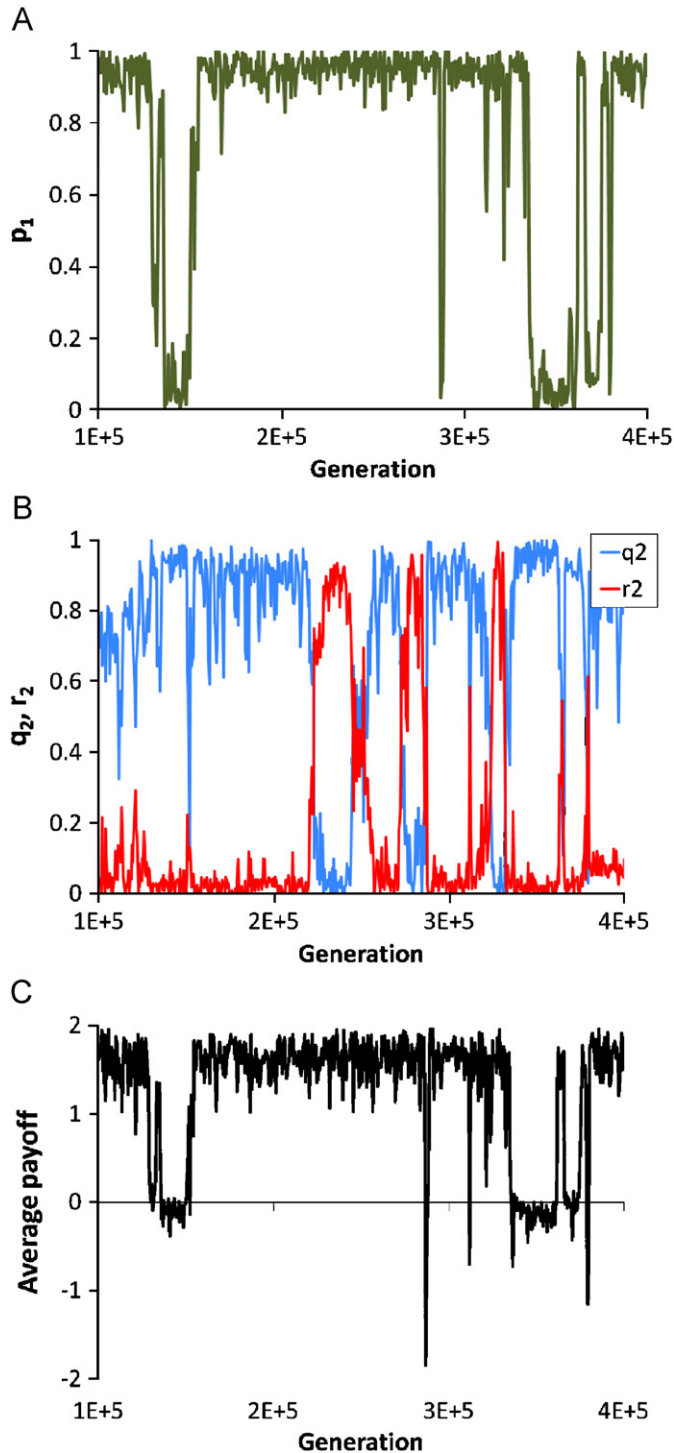


Fig. 2. The dynamics of finite populations. Representative simulation dynamics using payoff values $b = 3, c = 1, d = e = 1, \alpha = 1$, and $\beta = 4$ are shown, with timeseries of p_1 (A), q_2 (B), r_2 (B), and average payoff (C). The most time is spent in strategies near tit-for-tat ($p_1 \approx 1, q_2 \approx 1, r_2 \approx 0$). Sometimes a cooperative strategy using punishment arises ($p_1 \approx 1, q_2 < r_2$), or cooperation breaks down and a strategy near ALLD ($p_1 \approx 0, q_2 \approx 1, r_2 \approx 0$) becomes most common. Average payoff for tit-for-tat strategies is approximately equal to that of cooperative strategies that use punishment. This is a result of there being very little defection to respond to in both cases. Thus, here the response to defection does not greatly affect average payoffs of cooperative strategies.

The population is initialized with N players using a strategy close to ALLD, $p_i = 0.0001, q_i = 0.9998$ and $r_i = 0.0001$. Each simulation result discussed below is the average of four simulations each lasting 5×10^6 generations, for a total of 2×10^7 generations. Unless otherwise indicated, all simulations use the following parameter values: $N = 50, \mu = 0.1$, and $\tau = 0.8$.

Fig. 2 shows representative simulation dynamics. At any given time, most players in the population use the same strategy. New mutants arise frequently. Often mutants gain a foothold in the population, occasionally become more numerous, and then die out. Sometimes, a new mutant takes over the population. The process then repeats itself with novel mutations arising in a population where most players use this new resident strategy. Due to the stochastic nature of finite populations, less fit mutants sometimes go to fixation, and fitter mutants sometimes become extinct.

3.2. Comparison with Nash equilibrium analysis

3.2.1. Punishment use and relative cost of cooperation versus punishment

Our analysis in Section 2 found that for certain parameter values cooperative Nash equilibria exist, and that the best cooperative Nash equilibria may use punishment only if $\alpha \leq c$. If $\alpha > c$, punishment is never used and the response to (an occasional) P is C . We now examine whether similar results are found in our finite population simulations. We use the payoff values $b = 3, c = 1, d = e = 1$, and $\beta = 4$, and compare the dynamics of $\alpha = 0.1$ with $\alpha = 10$.

As shown in Fig. 3, the time average frequency of C, D , and P use are similar for both values of α . Most moves are cooperation ($\alpha = 0.1 : C = 78.7\%; \alpha = 10 : C = 87.6\%$). Defection occurs less frequently ($\alpha = 0.1 : D = 17.0\%; \alpha = 10 : D = 10.5\%$). Punishment is rare ($\alpha = 0.1 : P = 4.3\%; \alpha = 10 : P = 1.9\%$).

These simulation results are consistent with the Nash equilibrium analysis. The high level of cooperation observed in the simulations agrees with the presence of cooperative Nash equilibria for the chosen payoff values. The $\alpha > c$ simulation contained more C , less D , and less P than the $\alpha < c$ simulation. This also agrees with the Nash equilibrium analysis. If $\alpha > c$, the Nash equilibrium response to punishment is full cooperation, and

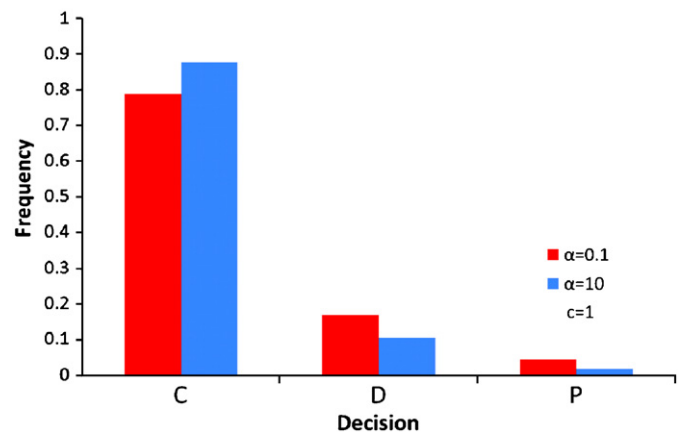


Fig. 3. The relative size of α and c has little effect on move frequencies. The time average frequency of cooperation, defection, and punishment are shown for $\alpha = 0.1$ and $\alpha = 10$, with $b = 3, c = 1, d = e = 1$, and $\beta = 4$. Simulation parameters $\mu = 0.1$, and $\tau = 0.8$ are used. Move use is time averaged over $N = 50$ players, playing for a total of 2×10^7 generations. Consistent with the Nash equilibrium analysis, there is a high level of cooperation in both cases, and the $\alpha > c$ simulation contains slightly more C , less D , and less P than the $\alpha < c$ simulation.

punishment is never used in response to defection. Therefore, you would expect to find more C, less D, and less P if $\alpha > c$ than if $\alpha < c$. The magnitude of these differences in move frequencies is small,

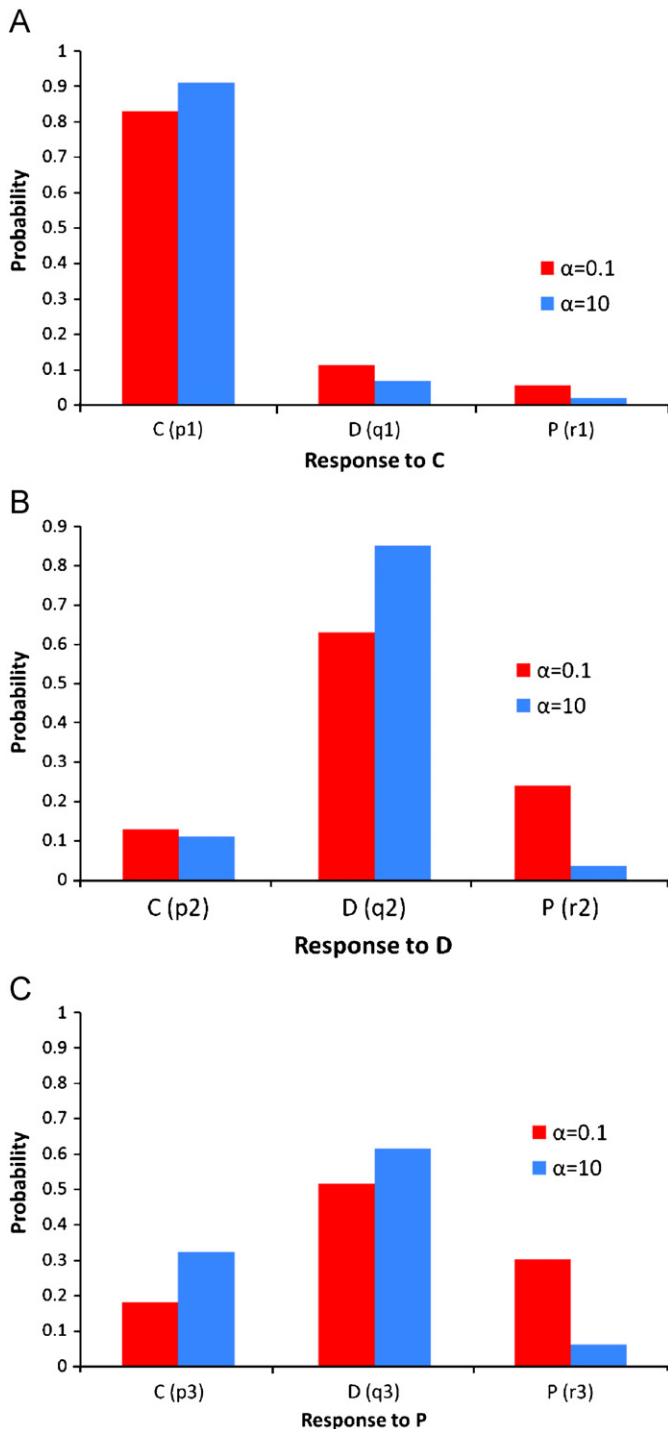


Fig. 4. As predicted by Nash equilibrium analysis, we see less punishment and more cooperation if $\alpha > c$ than if $\alpha < c$. Shown are strategy time averages for $\alpha = 0.1$ and $\alpha = 10$, with $b = 3, c = 1, d = e = 1$, and $\beta = 4$. Simulation parameters $\mu = 0.1$, and $\tau = 0.8$ are used. Strategies are time averaged over $N = 50$ players, playing for a total of 2×10^7 generations. There is agreement between the Nash equilibria analysis and the computer simulations on the high level of mutual cooperation regardless of the value of α , and the low level of punishment when $\alpha > c$. However, the computer simulations find that even when $\alpha < c$, the response to non-cooperation is much more likely to be defection than punishment. This suggests that of the infinitely many possible cooperative Nash equilibria, those that use tit-for-tat style defection in response to defection are favored by evolution over those that use costly punishment in response to defection.

which is also in agreement with the Nash equilibrium analysis: most of the time is spent in mutual cooperation, which is unaffected by the cost of punishment.

The time average strategies for both simulations are shown in Fig. 4. In Fig. 4A, we see little difference in the response to cooperation between α values. Cooperation is the overwhelming response to cooperation ($\alpha = 0.1 : p_1 = 0.83; \alpha = 10 : p_1 = 0.91$). This is consistent with the Nash equilibrium analysis for both values of α . The background level of P in response to C that exists when $\alpha < c$ decreases when $\alpha > c$ ($\alpha = 0.1 : r_1 = 0.06; \alpha = 10 : r_1 = 0.02$). It is intuitive that such illogical punishment is less favorable when punishment is very costly.

In Fig. 4B, we see some differences between the two simulations in their response to defection. When $\alpha < c$, D is most often responded to with D ($\alpha = 0.1 : q_2 = 0.63$), but D is also sometimes responded to with P ($\alpha = 0.1 : r_2 = 0.24$). When $\alpha > c$, D is responded to with P much less often ($\alpha = 10 : r_2 = 0.04$), and D in response to D is much more common ($\alpha = 10 : q_2 = 0.85$). The lack of P in response to D when $\alpha > c$ is consistent with the Nash equilibrium analysis. The significant preference for D over P when $\alpha < c$, however, is somewhat surprising. The Nash equilibrium results suggest that D and P can both be used when $\alpha < c$, and so we would not expect that q_2 is necessarily much greater than r_2 . Additionally, we see much less forgiveness (C in response to D) in both simulations than predicted by the Nash equilibrium analysis ($\alpha = 0.1 : p_2 = 0.13; \alpha = 10 : p_2 = 0.11$). Presumably this is a consequence of increased randomness in small (finite) populations. A very small increase in generosity, p_2 , destabilizes GTFT by allowing the invasion of ALLD. Thus in finite populations, stable cooperative strategies must stay far from the GTFT generosity threshold, and therefore forgive defection less often than GTFT.

In Fig. 4C, we see major differences between the two simulations in their response to punishment. In both simulations, the most frequent response to P is D ($\alpha = 0.1 : q_3 = 0.52; \alpha = 10 : q_3 = 0.61$). However, the use of C and P are starkly different depending on the value of α . When $\alpha < c$, C is much less common than P ($\alpha = 0.1 : p_3 = 0.18, r_3 = 0.30$). When $\alpha > c$, the opposite is true: C is common ($\alpha = 10 : p_3 = 0.32$) whilst P is rare ($\alpha = 10 : r_3 = 0.06$). This preference for C in response to P instead of P in response to P when $\alpha > c$ is consistent with the Nash equilibrium analysis. The mix of D and P in response to P when $\alpha < c$ is also consistent with the Nash equilibrium analysis. Again, however, we see less forgiveness in both simulations than predicted by the Nash equilibrium analysis.

In summary, we find agreement between the Nash equilibrium analysis and the computer simulations on the high level of mutual cooperation regardless of the value of α , and the low level of punishment when $\alpha > c$. However, the computer simulations find that even when $\alpha < c$, the response to defection is much more likely to be defection than punishment. This suggests that of the infinitely many possible cooperative Nash equilibria, those that use tit-for-tat style defection in response to defection are favored by evolutionary dynamics in finite populations over those that use costly punishment in response to defection.

3.2.2. Does punishment promote the evolution of cooperation?

Our analysis in Section 2 found that if Eqs. (20) and (21) are satisfied, then costly punishment allows for cooperative Nash equilibria even if none would exist using only cooperation and defection. We now ask what strategies are selected in our finite population simulations as cooperation becomes less beneficial. We use the payoff values $c = 1, d = e = 1, \alpha = 1$, and $\beta = 4$, and examine the level of cooperation as b varies. Our computer simulations use $w = 1$ and $\varepsilon \rightarrow 0$. Therefore, Eqs. (20) and (21) are satisfied when $b < 1$.

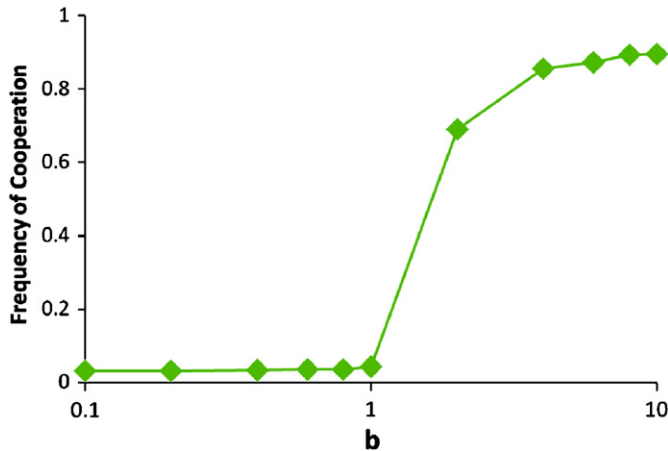


Fig. 5. Costly punishment does not promote the evolution of cooperation. Frequency of cooperation is shown as b is varied, with $c = 1, \alpha = 1, d = e = 1$, and $\beta = 4$. Simulation parameters $\mu = 0.1$, and $\tau = 0.8$ are used. Cooperation frequency is time averaged over $N = 50$ players, playing for a total of 2×10^7 generations. Cooperation is high when $b > 1$, and very low ($< 5\%$) when $b \leq 1$. This is the same pattern as would be seen in finite population simulations of classical direct reciprocity. Cooperation only succeeds where it would have in classical direct reciprocity. So we see that contrary to the Nash equilibrium analysis, costly punishment does not promote the evolution of cooperation in the framework of direct reciprocity.

As shown in Fig. 5, cooperation is the prevailing outcome when $b > 1$, whereas cooperation is rare ($< 5\%$) when $b \leq 1$. In the case that $b > 1$, cooperative Nash equilibria exist regardless of whether costly punishment is used. Thus it is not surprising to find a high level of cooperation in this parameter region. However, our Nash equilibrium analysis suggests that costly punishment could stabilize cooperation if $b < 1$. This does not seem to be the case in the finite population simulations. Contrary to the predictions of the Nash equilibrium analysis, costly punishment does not allow cooperation to evolve unless direct reciprocity alone would also be sufficient. Thus, in the context of direct reciprocity, punishment does not promote the evolution of cooperation.

3.3. Robustness to parameter variation

Our simulations suggest that for both $\alpha < c$ and $\alpha > c$, costly punishment is used less often than defection as a response to defection. Now we test how robust this finding is to variation in the payoff and simulation parameters. We choose the following baseline parameter values: $b = 3, c = 1, d = e = 1, \alpha = 1, \beta = 4, \tau = 0.8$, and $\mu = 0.1$. Each parameter is then varied, and for each value four simulations are run, each lasting 5×10^6 generations (for a total of 2×10^7 generations). The value of all strategy parameters p_i, q_i, r_i for each player are time averaged over the entire simulation time. As we are interested in the response to defection of essentially cooperative strategies (i.e. strategies that usually cooperate when playing against themselves), we only examine players with $p_1 > 0.75$. Among these cooperative players, we then examine the time average probabilities to defect in response to defection, q_2 , and to punish in response to defection, r_2 . Fig. 6 shows the results of varying payoffs β, α, d , and b , as well as simulation parameters τ, μ , and w on the time average values of q_2 and r_2 among cooperative players.

In Fig. 6A, we see that increasing β decreases defection use relative to punishment use. As punishment becomes more costly for the player who is punished, it becomes more effective to use P . Yet even with a 25 : 1 punishment technology, we find that $q_2 > r_2$.

In Fig. 6B, we see that increasing d (and e , as we assume that $d = e$) increases defection use relative to punishment use. As defection becomes more effective, it makes even less sense to punish. For $d > 4$, defection is more damaging to the recipient than punishment, while at the same time it is not costly to use to defect. Therefore, the probability to use defection in response to defection q_2 approaches 1. On the other extreme, even when defection is passive, $d = e = 0$, it is still true that $q_2 > r_2$.

In Fig. 6C, we see that increasing α increases defection use relative to punishment use. As punishment gets more expensive for the punisher, it becomes less effective to punish. Yet even if $\alpha = 0$, defection is still used more than punishment.

In Fig. 6D, we see that increasing b increases punishment use relative to defection. However, the probability to punish in response to defection r_2 never rises above $\frac{1}{3}$. As b increases, cooperation becomes increasingly prevalent and so there is less defection to respond to. When $b = 2$, we find that 28% of moves are D , as opposed to 4% D when $b = 25$. This reduces the selection pressure on players' response to defection, and both q_2 and r_2 approach chance, $\frac{1}{3}$. This has the effect of decreasing q_2 and increasing r_2 , but still it is always true that $q_2 > r_2$.

In Fig. 6E, we see that increasing the temperature of selection τ reduces selection intensity, and all move probabilities approach $\frac{1}{3}$. However, for all values of τ , it is never true that $q_2 < r_2$. Punishment is never favored over defection as a response to an opponent's defection.

In Fig. 6F, we see a similar pattern for increasing the mutation rate μ . As μ approaches 1, mutation dominates selection and all move probabilities approach $\frac{1}{3}$. But for all values $\mu < 1$, it is true that $q_2 > r_2$. Again, an opponent's D is always more often responded to with defection than with punishment.

In Fig. 6G, we relax the assumption that games are infinitely repeated. To make direct comparisons between $w = 1$ and $w < 1$ simulations possible, τ is increased by a factor of $1/(1 - w)$ when $w < 1$. This compensates for $w = 1$ payoffs reflecting average payoff per round, whereas $w < 1$ payoffs reflect total payoff. We see that for values $w < 1$, it is still true that $q_2 > r_2$. Thus the observation that defection is favored over punishment as a response to the opponent's defection applies to finite as well as infinitely repeated games.

The most striking aspect of Fig. 6 is that in all cases, $q_2 > r_2$ holds. The evolutionary preference for 'D in response to D' over 'P in response to D' is stable against variation in all parameter values. Hence we conclude that for reasonable parameter values, defection is always used more often than costly punishment as a response to defection. Evolution in finite populations disfavors strategies that use costly punishment in repeated games.

4. Comparison with experimental results

Many behavioral experiments have investigated the effect of costly punishment on human cooperation (Yamagishi, 1986; Ostrom et al., 1992; Fehr and Gächter, 2000, 2002; Page et al., 2005; Bochet et al., 2006; Güreker et al., 2006; Rockenbach and Milinski, 2006; Denant-Boemont et al., 2007; Dreber et al., 2008; Herrmann et al., 2008). We would like to compare our model predictions with the observed behavior in such experiments. However, the experimental setup used in most previous punishment studies differs from the situation described in this paper. The option to punish is offered as a separate stage following the decision to cooperate or defect. Only the design used by Dreber et al. (2008) is directly comparable with our model. Subjects

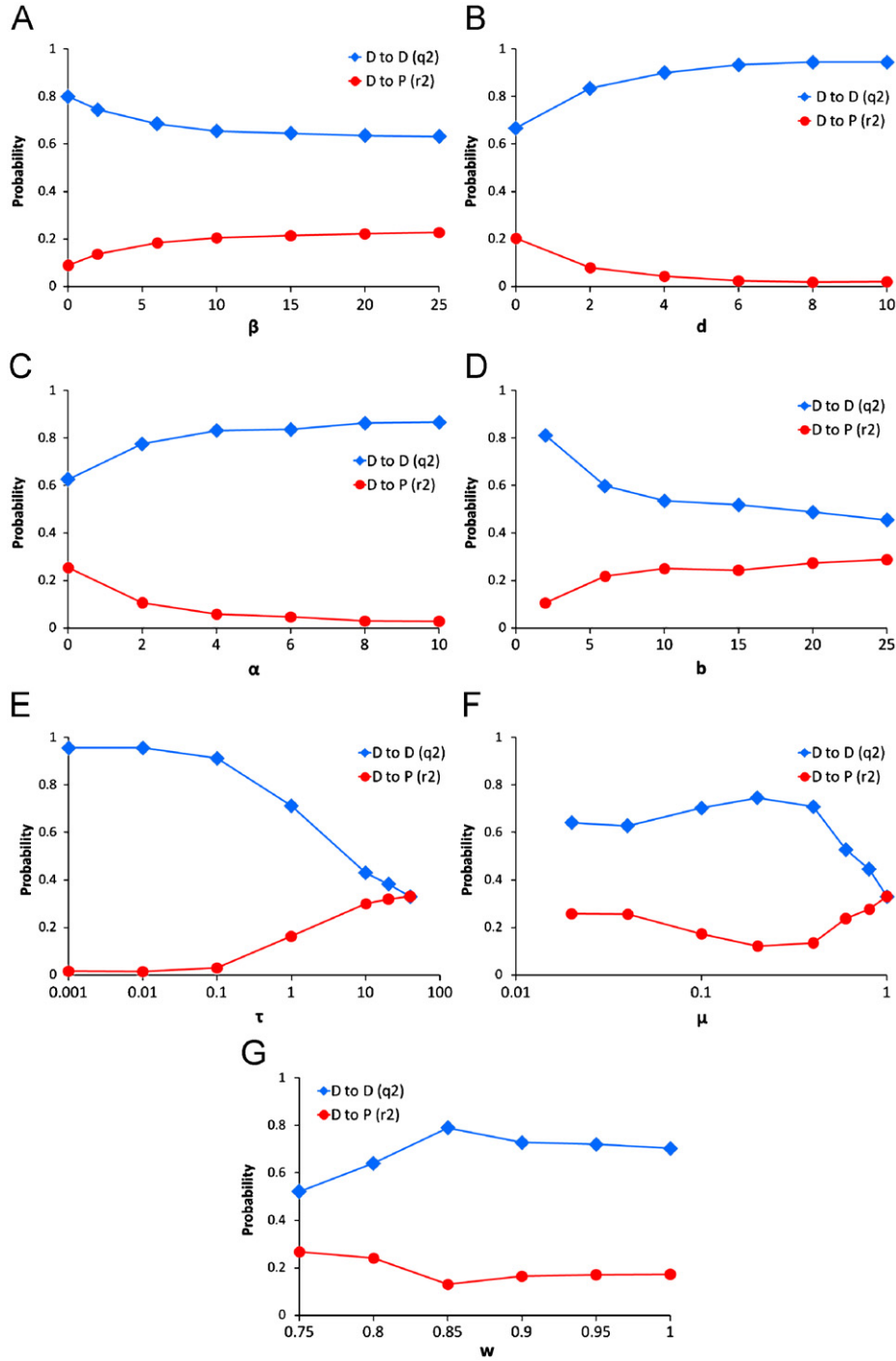


Fig. 6. Evolution disfavors the use of costly punishment across a wide range of payoff and simulation parameter values. Probabilities of D in response to D (q_2) and P in response to D (r_2) among cooperative strategies ($p_1 > 0.75$) are shown, time averaged over a total of 2×10^7 generations. Parameters β (A), d (B), α (C), b (D), τ (E), μ (F), and w (G) are varied. Values other than that which is varied are set to $b = 3, c = 1, d = e = 1, \alpha = 1, \beta = 4, \tau = 0.8, \mu = 0.1$, and $w = 1$. For all parameter sets explored, $q_2 > r_2$ holds: strategies that respond to D with D more often than P are selected by evolution. (A) As punishment becomes more costly for the player who is punished, it becomes more effective to use P . Yet even with a 25 : 1 punishment technology, we see $q_2 > r_2$. (B) As defection becomes more effective, it makes even less sense to punish. (C) As punishment gets more expensive for the punisher, it becomes less effective to punish. Yet even if $\alpha = 0$, defection is still used more than punishment. (D) Increasing b decreases the total number of D moves, and therefore decreases the selection pressure acting on q_2 and r_2 . This has the effect of moving both values towards $\frac{1}{3}$, thus decreasing q_2 and increasing r_2 . (E) Increasing the temperature of selection τ also reduces selection pressure, and all move probabilities approach $\frac{1}{3}$. Yet for all values of τ , we find $q_2 > r_2$. (F) As mutation rate μ increases, mutation dominates selection and all move probabilities approach $\frac{1}{3}$. Yet for all values $\mu < 1$, we find $q_2 > r_2$. (G) Even in finitely repeated games, $w < 1$, defection is favored over punishment, $q_2 > r_2$.

played a repeated 3-option Prisoner's Dilemma, with payoff structure as in Eq. (3). The payoff values $b = 3, c = 1, d = e = 1, \alpha = 1$, and $\beta = 4$ were used, and interactions had a continuation probability of $w = 0.75$. Given the similarity between this

experimental setup and the analysis presented here, we focus on the data from this experiment and make comparisons with predictions from (i) the Nash-equilibrium analysis and (ii) individual based simulations.

4.1. Nash equilibrium analysis

Following Dreber et al. (2008), we use the payoff values $b = 3$, $c = 1$, $d = e = 1$, $\alpha = 1$, and $\beta = 4$. It seems likely that human strategic instincts evolved in small societies where the probability

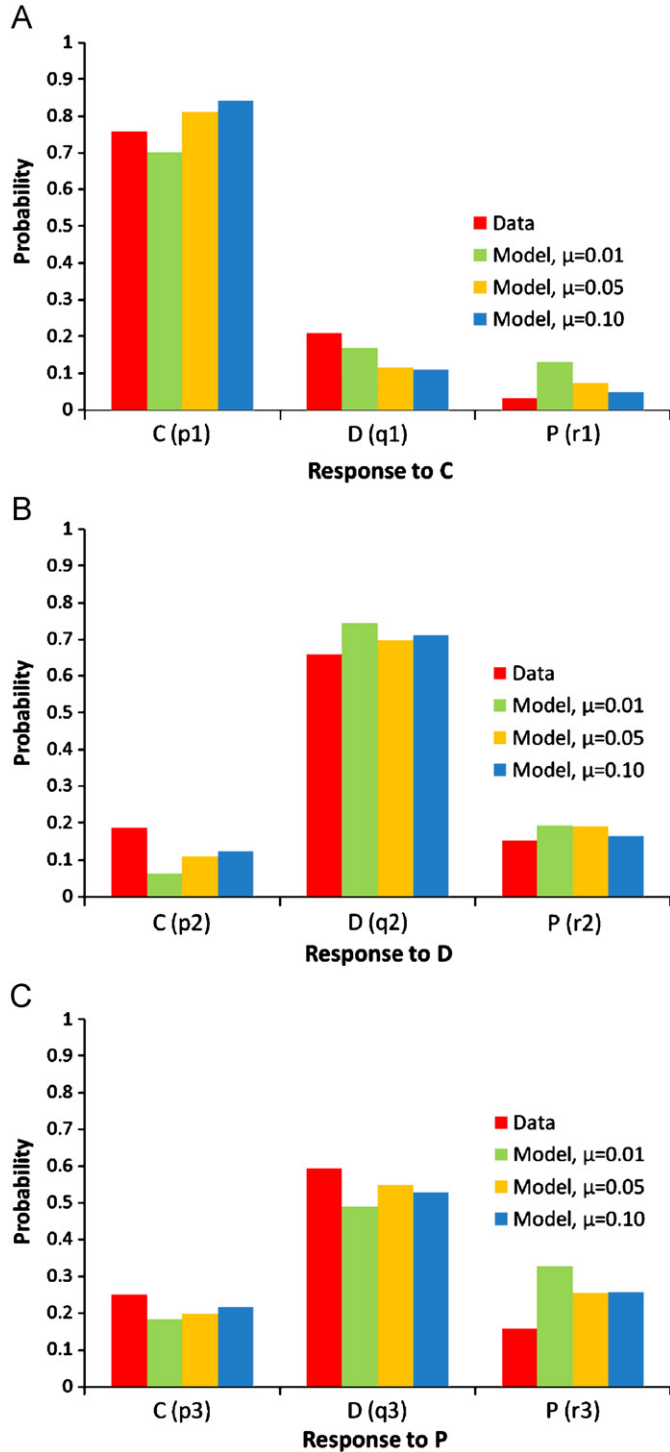


Fig. 7. We see quantitative agreement between the finite population simulations and human behavior in an experimental setting. Moreover, this agreement is robust to variation in the mutation rate μ . Time average strategies from Dreber et al. (2008) experimental data and finite population simulation using $\mu = 0.01$, $\mu = 0.05$, and $\mu = 0.1$ are shown. The simulation uses the Dreber et al. (2008) payoff values $b = 3, c = 1, d = e = 1, \alpha = 1$, and $\beta = 4$, and temperature $\tau = 0.8$. Such quantitative agreement demonstrates the power of finite population size analysis for describing human behavior.

of future interactions was always high. Thus, we study the limit $w \rightarrow 1$. According to our Nash equilibrium analysis in Section 2, we obtain

$$s = \begin{matrix} \text{Initial move} \\ \text{Response to C} \\ \text{Response to D} \\ \text{Response to P} \end{matrix} \begin{pmatrix} C & D & P \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ \frac{1}{2}\theta + \frac{5}{7}(1-\theta) & \frac{1}{2}\theta & \frac{2}{7}(1-\theta) \\ 1 & 0 & 0 \end{pmatrix} \quad (0 \leq \theta \leq 1), \tag{31}$$

as the best cooperative strategy in the limit of $w \rightarrow 1$ and $\varepsilon \rightarrow 0$ (see Fig. 1B). Against a defector, this strategy uses defection with probability 0.5 and no punishment, or uses punishment with probability 0.286 and no defection, or a linear combination of these values. Therefore, the possibility of using costly punishment is consistent with a cooperative Nash equilibrium. However, the experimental data suggest that the use of costly punishment is disadvantageous. More specifically, Dreber et al. (2008) found a strong negative correlation between total payoff and the probability to use P in response to an opponent's D . In this experimental situation, winners use a tit-for-tat like strategy while losers use costly punishment. Hence, Nash equilibrium analysis does not provide a good explanation of the observed experimental data.

4.2. Computer simulations

Consistent with the idea that humans evolved in settings where future encounters were likely, we find quantitative agreement between the experimental data and finite population simulations using $w = 1$ (Fig. 7). The optimal value of $\tau = 0.8$ was determined by minimizing the sum of squared differences between model predictions and observed data. As shown in Fig. 7, this fit is robust to variation in the mutation rate μ .

In both the computer simulations and the Nash equilibrium analysis, defection is used much more often than punishment after the opponent defects or punishes. We also see a similar use of cooperation: cooperation is reciprocated, but unlike in the Nash equilibrium analysis, the computer simulations find that it is uncommon to cooperate after the opponent has defected or punished. The agreement between computer simulations and experimental data demonstrates the power of finite population size analysis (Nowak et al., 2004; Imhof and Nowak, 2006) for characterizing human behavior, as does the robustness of the fit to variation in the mutation rate μ . With this method, ad hoc assumptions about other-regarding preferences are not needed to recover human behavior in the repeated Prisoner's Dilemma with costly punishment.

5. Discussion

It is important to study the effect of punishment on human behavior. Costly punishment is always a form of direct or indirect reciprocity. If I punish you because you have defected with me, then it is direct reciprocity. If I punish you because you have defected with others, then it is indirect reciprocity. Therefore, the precise approach for the study of costly punishment is to extend cooperation games from two possible moves, C and D , to three possible moves, C , D , and P and then study the consequences. In order to understand whether costly punishment can really promote cooperation, we must examine the interaction between costly punishment and direct or indirect reciprocity.

There are two essential questions to ask about such extended cooperation games. Should costly punishment be the response to a co-player's defection, instead of defection for defection as in

classical direct reciprocity? And does the addition of costly punishment allow cooperation to succeed in situations where direct or indirect reciprocity without costly punishment do not? Here we have explored these questions for the set of reactive strategies in the framework of direct reciprocity.

We have calculated all Nash equilibria among these reactive strategies. A subset of those Nash equilibria are cooperative, which means that these strategies would always cooperate with each other in the absence of errors. We find that if the cost of cooperation, c , is less than the cost of punishment, α , then the only cooperative Nash equilibrium is generous-tit-for-tat, which does not use costly punishment. However if the cost of cooperation, c , is greater than the cost of punishment, α , then there are infinitely many cooperative Nash equilibria and the response to a co-player's defection can be a mixture of C , D and P . We also find that the option for costly punishment allows such cooperative Nash equilibria to exist in parameter regions where there would have been no cooperation in classical direct reciprocity (including the case where the cost of cooperation, c , is greater than the benefit of cooperation, b , making cooperation unprofitable). Therefore if $\alpha < c$, it is possible for a cooperative Nash equilibrium to use costly punishment.

We have also performed computer simulations to study how evolutionary dynamics in finite sized population choose among all strategies in our strategy space. We find that for all parameter choices that we have investigated, costly punishment, P , is used less often than defection, D , in response to a co-player's defection. In these simulations, we also find that costly punishment fails to stabilize cooperation when cost of cooperation, c , is greater than the benefit of cooperation, b . Therefore, in the context of repeated interactions (1) natural selection opposes the use of costly punishment, and (2) costly punishment does not promote the evolution of cooperation. Winning strategies tend to stick with generous-tit-for-tat and ignore costly punishment, even if the cost of punishment, α , is less than the cost of cooperation, c .

The results of our computer simulations are in quantitative agreement with data from an experimental study (Dreber et al., 2008). In this game, people behave as predicted from the calculations of evolutionary dynamics, as opposed to the Nash equilibrium analysis. This agreement supports the validity of the idea that analysis of evolutionary dynamics in populations of finite size helps to understand human behavior.

Perhaps the existence of cooperative Nash equilibria that use costly punishment lies behind some people's intuition about costly punishment promoting cooperation. But our evolutionary simulations indicate that these strategies are not selected in repeated games.

In summary, we conclude that in the framework of direct reciprocity, selection does not favor strategies that use costly punishment, and costly punishment does not promote the evolution of cooperation.

Acknowledgments

Support from the John Templeton Foundation, the NSF/NIH joint program in mathematical biology (NIH grant R01GM078986), and J. Epstein is gratefully acknowledged.

Appendix A

In this appendix we will analytically derive all Nash equilibria of the repeated game.

We notice that saying strategy s is a Nash equilibrium is equivalent to saying that s is a best response to s . In other words, s

is a Nash equilibrium if it is a strategy that maximizes the payoff function

$$u(\cdot, s). \tag{A.1}$$

For each strategy s , we ask what strategy s^* is the best response to s . If s^* happens to be the same as s , then s is a Nash equilibrium.

For that purpose, we shall use a technique in dynamic optimization (Bellman, 1957). Let us define the 'value' of cooperation, defection, and punishment when the opponent uses strategy s . The value of each action is defined as the sum of its immediate effect and its future effect. For example, if you cooperate with an s -opponent, you immediately lose the cost of cooperation payoff, c . In the next round (which exists with probability w), however, the s -opponent reacts to your cooperation with cooperation, defection, or punishment, with probabilities p_1, q_1 and r_1 , respectively. Because we consider reactive strategies, your cooperation in round t has no effects on your payoff in round $t + 2$ or later. Thus the value of cooperation, v_C , is given by

$$v_C = \underbrace{-c}_{\text{immediate payoff}} + \underbrace{w(bp_1 - eq_1 - \beta r_1)}_{\text{future payoff}} \tag{A.2}$$

in the absence of errors. When the effect of errors is incorporated, the value of each action is given by

$$\begin{aligned} v_C &= -c + w(1 - 3\varepsilon)(bp_1 - eq_1 - \beta r_1) + w\varepsilon(b - e - \beta), \\ v_D &= +d + w(1 - 3\varepsilon)(bp_2 - eq_2 - \beta r_2) + w\varepsilon(b - e - \beta), \\ v_P &= -\alpha + w(1 - 3\varepsilon)(bp_3 - eq_3 - \beta r_3) + w\varepsilon(b - e - \beta). \end{aligned} \tag{A.3}$$

Given Eq. (A.3), the best response to strategy s is surprisingly simple: it is 'always take the action whose value is the largest'. If there is more than one best action, then any combination of such best actions is a best responses.

Depending on the relative magnitudes of v_C, v_D and v_P , we obtain seven different cases.

A.1. When $v_C > v_D, v_P$

The best response to s is 'always cooperate';

$$s^* = \begin{matrix} \text{Initial move} \\ \text{Response to } C \\ \text{Response to } D \\ \text{Response to } P \end{matrix} \begin{pmatrix} C & D & P \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}. \tag{A.4}$$

Let us assume that $s^* = s$ holds. Substituting s^* for s in Eq. (A.3) gives us

$$\begin{aligned} v_C &= -c + w(1 - 3\varepsilon)b, \\ v_D &= +d + w(1 - 3\varepsilon)b, \\ v_P &= -\alpha + w(1 - 3\varepsilon)b. \end{aligned} \tag{A.5}$$

(note that we will neglect the common term $w\varepsilon(b - e - \beta)$ in v 's in the following). We see that v_C can never be the largest of the three. Contradiction. Therefore, there are no Nash equilibria in this case.

A.2. When $v_D > v_P, v_C$

The best response to s is 'always defect';

$$s^* = \begin{matrix} \text{Initial move} \\ \text{Response to } C \\ \text{Response to } D \\ \text{Response to } P \end{matrix} \begin{pmatrix} C & D & P \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}. \tag{A.6}$$

Let us assume that $s^* = s$ holds. Substituting s^* for s in Eq. (A.3) gives us

$$\begin{aligned} v_C &= -c + w(1 - 3\varepsilon)(-e), \\ v_D &= +d + w(1 - 3\varepsilon)(-e), \\ v_P &= -\alpha + w(1 - 3\varepsilon)(-e). \end{aligned} \quad (\text{A.7})$$

Hence v_D is always the largest, which is consistent with our previous assumption. Therefore, Eq. (A.6) is always a Nash-equilibrium strategy.

A.3. When $v_P > v_C, v_D$

The best response to s is ‘always punish’;

$$s^* = \begin{matrix} \text{Initial move} \\ \text{Response to C} \\ \text{Response to D} \\ \text{Response to P} \end{matrix} \begin{pmatrix} C & D & P \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}. \quad (\text{A.8})$$

Let us assume that $s^* = s$ holds. Substituting s^* for s in Eq. (A.3) gives us

$$\begin{aligned} v_C &= -c + w(1 - 3\varepsilon)(-\beta), \\ v_D &= +d + w(1 - 3\varepsilon)(-\beta), \\ v_P &= -\alpha + w(1 - 3\varepsilon)(-\beta). \end{aligned} \quad (\text{A.9})$$

Obviously, v_D is the largest of the three. Contradiction. Therefore, there are no Nash equilibria in this case.

A.4. When $v_D = v_P > v_C$

The best response to s is ‘never cooperate’;

$$s^* = \begin{matrix} \text{Initial move} \\ \text{Response to C} \\ \text{Response to D} \\ \text{Response to P} \end{matrix} \begin{pmatrix} C & D & P \\ 0 & 1 - r_0 & r_0 \\ 0 & 1 - r_1 & r_1 \\ 0 & 1 - r_2 & r_2 \\ 0 & 1 - r_3 & r_3 \end{pmatrix}. \quad (\text{A.10})$$

Let us assume that $s^* = s$ holds. Substituting s^* for s in Eq. (A.3) gives us

$$\begin{aligned} v_C &= -c + w(1 - 3\varepsilon)\{(e - \beta)r_1 - e\}, \\ v_D &= +d + w(1 - 3\varepsilon)\{(e - \beta)r_2 - e\}, \\ v_P &= -\alpha + w(1 - 3\varepsilon)\{(e - \beta)r_3 - e\}. \end{aligned} \quad (\text{A.11})$$

The strategies described by Eq. (A.10) are Nash equilibria if the condition $v_D = v_P > v_C$ holds, which is equivalent to

$$\begin{aligned} c + d &> w(1 - 3\varepsilon)(e - \beta)(r_1 - r_2), \\ d + \alpha &= w(1 - 3\varepsilon)(e - \beta)(r_3 - r_2). \end{aligned} \quad (\text{A.12})$$

A.5. When $v_P = v_C > v_D$

The best response to s is ‘never defect’;

$$s^* = \begin{matrix} \text{Initial move} \\ \text{Response to C} \\ \text{Response to D} \\ \text{Response to P} \end{matrix} \begin{pmatrix} C & D & P \\ 1 - r_0 & 0 & r_0 \\ 1 - r_1 & 0 & r_1 \\ 1 - r_2 & 0 & r_2 \\ 1 - r_3 & 0 & r_3 \end{pmatrix}. \quad (\text{A.13})$$

Let us assume that $s^* = s$ holds. Substituting s^* for s in Eq. (A.3) gives us

$$\begin{aligned} v_C &= -c + w(1 - 3\varepsilon)\{-(b + \beta)r_1 + b\}, \\ v_D &= +d + w(1 - 3\varepsilon)\{-(b + \beta)r_2 + b\}, \\ v_P &= -\alpha + w(1 - 3\varepsilon)\{-(b + \beta)r_3 + b\}. \end{aligned} \quad (\text{A.14})$$

The strategies described by Eq. (A.13) are Nash equilibria if the condition $v_P = v_C > v_D$ holds, which is equivalent to

$$\begin{aligned} c + d &< w(1 - 3\varepsilon)(b + \beta)(r_2 - r_1), \\ c - \alpha &= w(1 - 3\varepsilon)(b + \beta)(r_3 - r_1). \end{aligned} \quad (\text{A.15})$$

A.6. When $v_C = v_D > v_P$

The best response to s is ‘never punish’;

$$s^* = \begin{matrix} \text{Initial move} \\ \text{Response to C} \\ \text{Response to D} \\ \text{Response to P} \end{matrix} \begin{pmatrix} C & D & P \\ 1 - q_0 & q_0 & 0 \\ 1 - q_1 & q_1 & 0 \\ 1 - q_2 & q_2 & 0 \\ 1 - q_3 & q_3 & 0 \end{pmatrix}. \quad (\text{A.16})$$

Let us assume that $s^* = s$ holds. Substituting s^* for s in Eq. (A.3) gives us

$$\begin{aligned} v_C &= -c + w(1 - 3\varepsilon)\{-(b + e)q_1 + b\}, \\ v_D &= +d + w(1 - 3\varepsilon)\{-(b + e)q_2 + b\}, \\ v_P &= -\alpha + w(1 - 3\varepsilon)\{-(b + e)q_3 + b\}. \end{aligned} \quad (\text{A.17})$$

The strategies described by Eq. (A.16) are Nash equilibria if the condition $v_C = v_D > v_P$ holds, which is equivalent to

$$\begin{aligned} c + d &= w(1 - 3\varepsilon)(b + e)(q_2 - q_1), \\ c - \alpha &< w(1 - 3\varepsilon)(b + e)(q_3 - q_1). \end{aligned} \quad (\text{A.18})$$

A.7. When $v_C = v_D = v_P$

Any strategy is a best response to s ;

$$s^* = \begin{matrix} \text{Initial move} \\ \text{Response to C} \\ \text{Response to D} \\ \text{Response to P} \end{matrix} \begin{pmatrix} C & D & P \\ p_0 & q_0 & r_0 \\ p_1 & q_1 & r_1 \\ p_2 & q_2 & r_2 \\ p_3 & q_3 & r_3 \end{pmatrix}. \quad (\text{A.19})$$

Let us assume that $s^* = s$ holds. For the three values of action, Eq. (A.3), to be the same, we need

$$\begin{aligned} -c + w(1 - 3\varepsilon)(bp_1 - eq_1 - \beta r_1) \\ = +d + w(1 - 3\varepsilon)(bp_2 - eq_2 - \beta r_2) \\ = -\alpha + w(1 - 3\varepsilon)(bp_3 - eq_3 - \beta r_3). \end{aligned} \quad (\text{A.20})$$

The strategies described by Eq. (A.19) are Nash equilibria if the condition Eq. (A.20) is satisfied.

Appendix B

B.1. Mutation kernel

When a learner switches strategies, with probability μ he experiments with a new strategy as opposed to adopting the strategy of the teacher. In this case, the probabilities p_i , q_i , and r_i

for each i ($i = 1, 2, 3$ for C, D, P) are randomly assigned as follows. Two random numbers X_1 and X_2 are drawn from a U-shaped beta distribution described by the probability density function

$$f(x) = \frac{x^{\gamma-1}(1-x)^{\gamma-1}}{\int_0^1 u^{\gamma-1}(1-u)^{\gamma-1} du}. \quad (\text{B.1})$$

The simulations presented here use $\gamma = 0.1$. Varying γ affects the total level of cooperation (p_1), but it is still the case that $q_2 > r_2$. Without loss of generality, let $X_1 < X_2$. Three random numbers between 0 and 1, X'_1, X'_2 , and X'_3 are then generated using X_1 and X_2 :

$$X'_1 = X_1, \quad (\text{B.2})$$

$$X'_2 = X_2 - X_1, \quad (\text{B.3})$$

$$X'_3 = 1 - X_2. \quad (\text{B.4})$$

Finally, p_i, q_i , and r_i are randomly matched with X'_1, X'_2 , and X'_3 . This order randomization is necessary to preserve $p_i + q_i + r_i = 1$ while still maintaining the same probability distribution for each variable. The resulting distribution is U-shaped, with the probability density near 0 twice as large as that near 1.

References

- Alexander, R.D., 1987. *The Biology of Moral Systems*. Aldine de Gruyter, New York.
- Axelrod, R., 1984. *The Evolution of Cooperation*. Basic Books, USA, New York.
- Axelrod, R., 1986. An evolutionary approach to norms. *Amer. Polit. Sci. Rev.* 80, 1095–1111.
- Axelrod, R., Hamilton, W.D., 1981. The evolution of cooperation. *Science* 211, 1390–1396.
- Bellman, R., 1957. *Dynamic Programming*. Princeton University Press, Princeton, NJ.
- Bochet, O., Page, T., Putterman, L., 2006. Communication and punishment in voluntary contribution experiments. *J. Econ. Behav. Org.* 60, 11–26.
- Boerlijst, M.C., Nowak, M.A., Sigmund, K., 1997. The logic of contrition. *J. Theor. Biol.* 185, 281–293.
- Bowles, S., 2001. Individual interactions, group conflicts, and the evolution of preferences. In: Durlauf, S.N., Young, H.P. (Eds.), *Social Dynamics*. MIT Press, Cambridge, MA, pp. 155–190.
- Boyd, R., Richerson, P., 1990. Group selection among alternative evolutionarily stable strategies. *J. Theor. Biol.* 145, 331–342.
- Boyd, R., Richerson, P.J., 1992. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol. Sociobiol.* 13, 171–195.
- Boyd, R., Gintis, H., Bowles, S., Richerson, P.J., 2003. Evolution of altruistic punishment. *Proc. Natl. Acad. Sci. USA* 100, 3531–3535.
- Brandt, H., Sigmund, K., 2006. The good, the bad and the discriminator—errors in direct and indirect reciprocity. *J. Theor. Biol.* 239, 183–194.
- Brandt, H., Hauert, C., Sigmund, K., 2003. Punishment and reputation in spatial public goods games. *Proc. R. Soc. Lond. B* 270, 1099–1104.
- Burnham, T.C., Johnson, D.P., 2005. The biological and evolutionary logic of human cooperation. *Analyse und Kritik* 27, 113–135.
- Carlsmith, K., Darley, J.M., Robinson, P.H., 2002. Why do we punish? Deterrence and just deserts as motives for punishment. *J. Pers. Soc. Psychol.* 83, 284–299.
- Clutton-Brock, T.H., Parker, G.A., 1995. Punishment in animal societies. *Nature* 373, 209–216.
- Denant-Boemont, L., Masclet, D., Noussair, C.N., 2007. Punishment, counter-punishment and sanction enforcement in a social dilemma experiment. *Econ. Theory* 33, 145–167.
- Dreber, A., Rand, D.G., Fudenberg, D., Nowak, M.A., 2008. Winners don't punish. *Nature* 452, 348–351.
- Fehr, E., Gächter, S., 2000. Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* 90, 980–994.
- Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. *Nature* 415, 137–140.
- Fowler, J.H., 2005. Altruistic punishment and the origin of cooperation. *Proc. Natl. Acad. Sci.* 102, 7047–7049.
- Fudenberg, D., Maskin, E., 1990. Evolution and cooperation in noisy repeated games. *Am. Econ. Rev.* 80, 274–279.
- Fudenberg, D., Tirole, J., 1991. *Game Theory*. MIT Press, Cambridge, MA.
- Gürerk, O., Irlenbusch, B., Rockenbach, B., 2006. The competitive advantage of sanctioning institutions. *Science* 312, 108–111.
- Hauert, C., Traulsen, A., Brandt, H., Nowak, M.A., Sigmund, K., 2007. Via freedom to coercion: the emergence of costly punishment. *Science* 316, 1905–1907.
- Herrmann, B., Thöni, C., Gächter, S., 2008. Antisocial punishment across societies. *Science* 319, 1362–1367.
- Imhof, L.A., Nowak, M.A., 2006. Evolutionary game dynamics in a Wright–Fisher process. *J. Math. Biol.* 52, 667–681.
- Imhof, L.A., Fudenberg, D., Nowak, M.A., 2005. Evolutionary cycles of cooperation and defection. *Proc. Natl. Acad. Sci. USA* 102, 10797–10800.
- Imhof, L.A., Fudenberg, D., Nowak, M.A., 2007. Tit-for-tat or win–stay, lose–shift? *J. Theor. Biol.* 247, 574–580.
- Kahneman, D., Schkade, D., Sunstein, C.R., 1998. Shared outrage and erratic awards: the psychology of punitive damages. *J. Risk Uncertainty* 16, 49–86.
- Kandori, M., 1992. Social norms and community enforcement. *Rev. Econ. Stud.* 59, 63–80.
- Kraines, D., Kraines, V., 1989. Pavlov and the prisoner's dilemma. *Theory and Decision* 26, 47–79.
- Lindgren, K., 1991. Evolutionary phenomena in simple dynamics. In: Langton, C., et al. (Eds.), *Artificial Life II*. Addison-Wesley, Redwood City, CA, pp. 295–312.
- Lindgren, K., Nordahl, M.G., 1994. Evolutionary dynamics of spatial games. *Physica D* 75, 292–309.
- Lotem, A., Fishman, M.A., Stone, L., 1999. Evolution of cooperation between individuals. *Nature* 400, 226–227.
- May, R.M., 1987. More evolution of cooperation. *Nature* 327, 15–17.
- Molander, P., 1985. The optimal level of generosity in a selfish, uncertain environment. *J. Conflict Resolution* 29, 611–618.
- Nakamaru, M., Iwasa, Y., 2006. The coevolution of altruism and punishment: role of the selfish punisher. *J. Theor. Biol.* 240, 475–488.
- Nowak, M.A., 1990. Stochastic strategies in the prisoners dilemma. *Theor. Popul. Biol.* 38, 93–112.
- Nowak, M.A., 2006. Five rules for the evolution of cooperation. *Science* 314, 1560–1563.
- Nowak, M.A., Sigmund, K., 1989. Oscillations in the evolution of reciprocity. *J. Theor. Biol.* 137, 21–26.
- Nowak, M.A., Sigmund, K., 1990. The evolution of stochastic strategies in the Prisoner's Dilemma. *Acta Appl. Math.* 20, 247–265.
- Nowak, M.A., Sigmund, K., 1992. Tit for tat in heterogeneous populations. *Nature* 355, 250–253.
- Nowak, M.A., Sigmund, K., 1993. A strategy of win–stay, lose–shift that outperforms tit for tat in Prisoner's Dilemma. *Nature* 364, 56–58.
- Nowak, M.A., Sigmund, K., 1998. Evolution of indirect reciprocity by image scoring. *Nature* 393, 573–577.
- Nowak, M.A., Sigmund, K., 2005. Evolution of indirect reciprocity. *Nature* 437, 1291–1298.
- Nowak, M.A., Sasaki, A., Taylor, C., Fudenberg, D., 2004. Emergence of cooperation and evolutionary stability in finite populations. *Nature* 428, 646–650.
- Ohtsuki, H., Iwasa, Y., 2004. How should we define goodness? Reputation dynamics in indirect reciprocity. *J. Theor. Biol.* 231, 107–120.
- Ohtsuki, H., Iwasa, Y., 2005. The leading eight: social norms that can maintain cooperation by indirect reciprocity. *J. Theor. Biol.* 239, 435–444.
- Ostrom, E., Walker, J., Gardner, R., 1992. Covenants with and without a sword: self-governance is possible. *Am. Pol. Sci. Rev.* 86, 404–417.
- Ostrom, E., Gardner, J., Walker, R., 1994. *Rules, Games, and Common-Pool Resources*. University of Michigan Press, Ann Arbor.
- Pacheco, J.M., Santors, F.C., Chalub, F.A.C.C., 2006a. Stern-judging: a simple, successful norm which promotes cooperation under indirect reciprocity. *PLoS Comput. Biol.* 2, 1634–1638.
- Pacheco, J.M., Traulsen, A., Nowak, M.A., 2006b. Active linking in evolutionary games. *J. Theor. Biol.* 243, 437–443.
- Page, T., Putterman, L., Unel, B., 2005. Voluntary association in public goods experiments: reciprocity mimicry and efficiency. *Econ. J.* 115, 1032–1053.
- Panchanathan, K., Boyd, R., 2005. A tale of two defectors: the importance of standing for evolution of indirect reciprocity. *J. Theor. Biol.* 224, 115–126.
- Rockenbach, B., Milinski, M., 2006. The efficient interaction of indirect reciprocity and costly punishment. *Nature* 444, 718–723.
- Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., Cohen, J.D., 2003. The neural basis of economic decision-making in the Ultimatum Game. *Science* 300, 1755–1758.
- Selten, R., Hammerstein, P., 1984. Gaps in Harley's argument on evolutionarily stable learning rules and in the logic of tit for tat. *Behav. Brain Sci.* 7, 115–116.
- Sigmund, K., 2008. Punish or perish: enforcement and reciprocation in human collaboration. *Trends Ecol. Evol.* 22, 593–600.
- Sigmund, K., Hauert, C., Nowak, M.A., 2001. Reward and punishment. *Proc. Natl. Acad. Sci. USA* 98, 10757–10762.
- Sugden, R., 1986. *The Economics of Rights, Cooperation and Welfare*. Blackwell, Oxford.
- Traulsen, A., Nowak, M.A., 2006. Evolution of cooperation by multilevel selection. *Proc. Natl. Acad. Sci. USA* 103, 10952–10955.
- Traulsen, A., Nowak, M.A., Pacheco, J.M., 2006. Stochastic dynamics of invasion and fixation. *Phys. Rev. E* 74, 011909.
- Traulsen, A., Pacheco, J.M., Nowak, M.A., 2007. Pairwise comparison and selection temperature in evolutionary game dynamics. *J. Theor. Biol.* 246, 522–529.
- Trivers, R.L., 1971. The evolution of reciprocal altruism. *Q. Rev. Biol.* 46, 35–57.
- Wahl, L.M., Nowak, M.A., 1999a. The continuous prisoner's dilemma: I. Linear reactive strategies. *J. Theor. Biol.* 200, 307–321.
- Wahl, L.M., Nowak, M.A., 1999b. The continuous prisoner's dilemma: I. Linear reactive strategies with noise. *J. Theor. Biol.* 200, 307–321.
- Wilson, D.S., 1975. A theory of group selection. *Proc. Natl. Acad. Sci.* 72, 143–146.
- Wilson, D.S., Sober, E., 1994. Reintroducing group selection to the human behavioral sciences. *Behav. Brain Sci.* 17, 585–654.
- Wynne-Edwards, V.C., 1962. *Animal Dispersion; in Relation to Social Behaviour*. Oliver and Boyd, London.
- Yamagishi, T., 1986. The provision of a sanctioning system as a public good. *J. Pers. Soc. Psychol.* 51, 110–116.
- Yamagishi, T., 1988. Seriousness of social dilemmas and the provision of a sanctioning system. *Soc. Psychol. Q.* 51, 32–42.