Irenaeus Wolff

Retaliation and the role for punishment in the evolution of cooperation

Member of

thurgauwissenschaft

www.thurgau-wissenschaft.ch

**THURGAU INSTITUTE
OF ECONOMICS**
at the University of Konstanz

Retaliation and the role for punishment in the evolution of cooperation

Irenaeus Wolff

TWI / University of Konstanz, Hauptstrasse 90, 8280 Kreuzlingen, Switzerland
wolff@twi-kreuzlingen.ch

Abstract:

Models of evolutionary game theory have shown that punishment may be an adaptive behaviour in environments characterised by a social-dilemma situation. Experimental evidence closely corresponds to this finding but questions the cooperation-enhancing effect of punishment if players are allowed to retaliate against their punishers. This study provides a theoretical explanation for the existence of retaliating behaviour in the context of repeated social dilemmas and analyses the role punishment can play in the evolution of cooperation under these conditions. We show a punishing strategy can pave the way for a partially-cooperative equilibrium of conditional cooperators and defecting types and, under positive mutation rates, foster the cooperation level in this equilibrium by prompting reluctant cooperators to cooperate. However, when rare mutations occur, it cannot sustain cooperation by itself as punishment costs favour the spread of non-punishing cooperators.

Keywords: Public goods, Prisoner's Dilemma, Strong reciprocity, Counter-punishment *JEL:* C73, C72, H41

1 Introduction

Recent laboratory experiments have cast serious doubt on the scope of earlier findings suggesting that punishment is a suitable solution for social-dilemma situations. In particular, the very cooperation-enhancing effect has been challenged: by relaxing the restriction to one round of punishment and allowing for retaliation, Denant-Boemont et al. (2007) and Nikiforakis (2008) demonstrate how sensitive the cooperative outcome of earlier studies is to changes in the experimental setup (e.g., Yamagishi, 1986, or Fehr and Gächter, 2000).

1 INTRODUCTION

While the (exogenous) restriction to a single stage of punishment does not seem to be plausible in most real-world situations, existing (evolutionary) models of cooperation have comfortably rested on this assumption. The present paper makes a first step in addressing the challenge posed by the experimental results mentioned. In our model, we provide a possible explanation for why a retaliative strategy may be adaptive within the context of an iterated social-dilemma game: if punishers stop sanctioning free-riding behaviour after retaliation and conditional cooperators do not require full cooperation to maintain their cooperative attitude, retaliating defectors can reap the benefits from conditional cooperators' cooperation free of punishment without having to cooperate themselves. Within this framework, we show that punishing cooperators still can play two roles: they can break up an equilibrium of omnilateral defection and pave the way for a stable polimorphic equilibrium comprised of conditional cooperators and a mix of defecting types; and they can increase the level of cooperation within a polimorphic equilibrium if mutation plays a role. What punishment cannot do, in the confines of our framework, is to maintain a full-cooperation equilibrium if there is even a tiny mutation rate.

In the past, a variety of mechanisms have been proposed to account for the high level of cooperation among humans, such as kinship (Hamilton, 1964), reciprocal altruism (Trivers, 1971, or Axelrod and Hamilton, 1981), costly signaling (Zahavi, 1975, or Gintis et al., 2001), indirect reciprocity and reputation (Alexander, 1979 and 1987, Nowak and Sigmund, 1998, or Ohtsuki, Iwasa, and Nowak, 2009), 'culture' (Cavalli-Sforza and Feldman, 1981, or Boyd and Richerson, 1985), norms (Sugden, 1986, or Sethi, 1996), group selection (Sober and Wilson, 1998), and strong reciprocity (Gintis, 2000, or Gintis et al., 2003).[1] The concept of strong reciprocity has received a lot of attention, as it is unique among the explanations proposed in the literature for the evolution of cooperation in that it can account for the substantial degree of cooperation observed in anonymous laboratory settings even when there is no repeated interaction between the same players. While the concept may have different readings, the general idea is that of a behavioural disposition "to sacrifice resources to bestow benefits on those who have bestowed benefits" and "to sacrifice resources to punish those who are not bestowing benefits in accordance with some social norm".[2] In a prisoner's dilemma or

---

[1] The studies mentioned are only meant to indicate important early contributions. For a more detailed picture of the literature, the interested reader is kindly referred to, e.g., the works collected in Hammerstein (2003).

[2] Fehr and Henrich (2003, p.57); for a discussion of the different readings of strong reciprocity depending on whether this is seen as a *behaviour* (or behavioural algorithm) or a *strategy*, cf. Stephens (2005).

a public-good game like the ones used in the aforementioned models on the evolution of cooperation, this translates into a cooperative action in settings without repeated interaction as well as with repeated interaction, unless—in the latter case—if the other player(s) is (are) in bad standing. Additionally, strong reciprocity stipulates the punishment of defectors whenever there are punishment stages and regardless of players' anonymity.

It has been shown time and again that unconditional cooperation cannot be stabilised easily in such settings; at the same time, costly punishment can be evolutionary stable and thus support cooperation in societies even when there is no inter-group conflict (Henrich and Boyd, 2001, Bowles and Gintis, 2004, Carpenter et al., 2004, Carpenter, 2007). Recent contributions have gone on to show that strong reciprocators can even invade a non-cooperative society under certain conditions (Fowler, 2005, Hauert et al., 2007). However, virtually none of these studies has taken into account that in many real-life situations, punished players will have the opportunity to get back at their punishers. Rather, the standard view in the literature on the evolution of cooperation seems to be that "while free riders occasionally punish cooperators, they do so rarely enough that we [can] restrict the ability to punishment to cooperators."[3] If at all, models of cooperation have included additional punishment stages reserved to sanction enforcement (Henrich and Boyd, 2001).[4] The exception to the rule comes as a couple of papers proposed by Janssen and Bushman (2008) and Wolff (2009) who both simply assume retaliative behaviour into existence, relying on the experimental evidence to justify this assumption. However, in light of the fact that they do not explain why retaliation may arise in the first place (and that it cannot arise in the situations they study), it remains unclear what we can learn from these studies about those situations in which retaliative behaviour may be adaptive.

In a recent contribution, Rand et al. (2010) remove the restriction that punishment be targeted at defectors to question the punishment-based explanation to cooperation altogether. However, they do so by showing that, under a certain form of learning, punishing cooperators can be invaded by cooperators who punish everybody else—unconditionally or conditionally on making positive contributions—as these invaders obtain a locally higher payoff than those they punish, and behaviour is copied on the basis of local payoff comparisons. In their model, only defection coupled with unconditional punishment of everybody else is evolutionarily stable. However, in evolutionary terms, their model suggests living in communities is not adaptive, as there

---

[3] Carpenter et al. (2004, p. 409), addition by this author.

[4] Recent laboratory evidence suggests additional punishment stages are used for additional punishment assignments and retaliation, but rarely for sanction enforcement, e.g., Nicklisch and Wolff (2011).

are no gains from cooperation and agents merely destroy their own resources to destroy even more of others' resources. Note that it is the mere *possibility* of punishing others—which in most relevant situations of life in groups will be given—rather than the use of punishment to ensure cooperation that leads to the gloomy conclusion. The present paper proposes a different argument, suggesting that retaliation best be considered in the framework of repeated interactions: retaliation may have a purpose in the pursuit of resources. If group members are generally willing to cooperate even if a tiny fraction of the group defects, and if punishers cease to punish when retaliated against, retaliators may have an evolutionary advantage when interactions are long enough.

2 The game

At each moment in time, a large number of groups each consisting of N individuals are randomly drawn from a very large population. Groups remain constant for the duration of an indefinitely-repeated game with continuation probability π. The stage game consists of three stages. In stage 0, the contribution stage, agents play a symmetric N-person binary public-good game. A cooperating player incurs a net cost of c to convey a benefit of $b/(N-1)$ to every other member of the group, where $0 < c < b$, while a defecting player does not incur any cost nor convey any benefit. In the second stage, players may punish each other, incurring a cost of k and causing a damage of p for the player punished. In the third stage, finally, punished players may retaliate with the same $k : p$ punishment technology. Reproduction, or learning, takes place between interactions and follows the standard discrete-time replicator dynamics (Taylor and Jonker, 1978), and therefore, the fraction x_i of each agent type i evolves according to

$$x_i(\tau + 1) = \frac{u_0 + u_i\left(\mathbf{x}(\tau)\right) \cdot x_i(\tau)}{u_0 + \sum_{\forall j} u_j\left(\mathbf{x}(\tau)\right) \cdot x_j(\tau)}, \tag{1}$$

where u_0 is each agent's baseline fitness and $u_i\left(\mathbf{x}(\tau)\right)$ is an i-agent's average additional fitness, or payoff, from the interaction given the type distribution in the population, $\mathbf{x}(\tau)$, at time τ.

2.1 Strategies under consideration

Clearly, there is an infinite number of possible strategies—and therefore, of potential agent-types—in a public-good game that is repeated infinitely (albeit with a given ending probability), and that allows both for punishment

of other agents after each contribution stage and for retaliation against such
punishment. Hence, we need to restrict our attention to a subset of all
possible strategies. The following list presents a reduced set of agent-type
families, from which we will subsequently 'recruit' the four types making up
our model population.

D never cooperates, never punishes, and does not react to punishment.
The traditional defector-type;

R never cooperates, always retaliates. The R-type is a direct implemen-
tation of the idea that punished defectors may retaliate, as ample ex-
perimental evidence suggests many do (e.g., Nikiforakis, 2008);

T_n a nice tit-for-tat type who always cooperates if at least n other players
cooperated in the preceding stage game, and defects, otherwise. T_n's
never punish and revert to cooperation when punished.[5] T_n-types were
introduced by Joshi (1987) and Boyd and Richerson (1988);

P always contributes and always punishes any non-contributing player.
Early treatments of punishment behaviour date back at least to Axelrod
(1986);

DT_n starts by defecting, switching to T_n only once punished. DT_n-types are
straight-forward generalisations of the *reluctant-cooperator* type intro-
duced by Boyd and Richerson (1992);

PT_n acts like a P-type unless retaliated against; then immediately switches
to behaving like a T_n-type (if retaliated against in stage game τ, the
next move is conditioned on others' behaviour in stage game τ, too);
PT_n-types (i) may be adaptive in the presence of R's—and more adap-
tive than P's, as argued below—and (ii) may be the reason for retali-
ating behaviour to appear, in the first place. In the spirit of Boyd and
Richerson's (1992) model, they also may be seen as a kind of 'semi-
timid' punisher type;

PD acts like a P-type unless retaliated against; then immediately switches
to behaving like a D-type; the discussion of PT_n's also applies to PD's.

---

[5]The reaction to punishment was added for completeness as it is not part of the original
models; it will be irrelevant in our main model.

2.2   Additional assumptions and type choice

In the following, we will make two additional assumptions. First, restrict the attention to interaction groups of $N = 3$ individuals, to simplify the game enough to draw the main conclusions. Second, assume that $\frac{(N-2)}{(N-1)}b > c$, meaning that cooperating types who cooperate in the presence of a single defecting player still derive a strictly positive benefit from the interaction. This may be seen as a strong assumption given the restriction to groups of 3; however, there is no obvious reason why the main results presented in this paper should not carry over qualitatively to larger groups, in which case our second assumption seems no longer unrealistic. As an indication, we sketch a model variant with large N in the appendix.

Similar to Boyd and Richerson (1992), we disregard pure defectors in favour of other types that start free-riding but subsequently react to punishment in either of two ways: by cooperating in the next stage game or by retaliating. In terms of the type taxonomy above, we replace D's by DT_n's, on the one hand, and R-types, on the other.[6] Furthermore, we discard the P-type as it is dominated by PT_0, as well as the unconditionally cooperating (if only after punishment) T_0 and DT_0 which are dominated by T_1 and DT_1, respectively.[7] We abstract from PT_0's for a similar reason, even though there is a specific matching in which the PT_0 does better than a PT_1, namely when matched with an DT_1 and a R-agent; in all other matchings, the PT_0 does only equally well or worse. Finally, we can safely disregard the PD-type as it is behaviourally equivalent to the PT_{N-1}-type under the strategy set considered.

We are now left with seven agent types, R's and three type pairs involving conditional cooperation somewhere in their strategy specification: T_1 and T_2, DT_1 and DT_2, and PT_1 and PT_2. To arrive, finally, at a tractable strategy set, let us eliminate three of the latter, one of each pair. For DT_n's, eliminating DT_2 seems like the natural choice, for a similar argument as presented for PT_0's above: under many matchings, DT_1 and DT_2 do equally well; under a number of matchings, DT_1 does better; and only when matched with a PT_2 and an R-agent, the DT_2 does better, while in a group with a PT_1 and an

---

[6]Boyd and Richerson (1992, p.177) justify this choice in two ways, by an "abundant empirical evidence that organisms do respond to punishment," and pointing out that this merely is a best case for the evolution of cooperation. We briefly re-introduce pure defectors in section 2.5 on equilibrium stability with respect to excluded types.

[7]Strictly speaking, all of the above dominance relationships referred to are weak, and hence, the corresponding strategies need not be driven to extinction. However, as long as we are focusing on interior solution paths, i.e., punishing respectively retaliating re-spectively free-riding types still occur, D's, P's, T_0's, and DT_0's are selected against more strongly than the corresponding DT_{N-1}-, PT_0-, T_1-, and DT_1-types.

R-agent, either type may do better, depending on the parameter choices (and the resolution of the coordination problem between the PT_1 and the DT_1). A sensible choice between the T_n- and the PT_n-types, respectively, is less obvious. The early works of Joshi (1987) and Boyd and Richerson (1988) have shown that in a world of only T_n-types competing with D's, the only stable conditional-cooperator type is T_{N-1} (given all other types can be exploited by a single defector). However, both go on to show that the conditions for T_{N-1} to invade a defecting population are extreme. Two characteristics of the present model suggest different aspects may be important in our case: most importantly, we incorporate punishment and defector types may react to being sanctioned; second, we have posited above that there is a mutual benefit from cooperation even if only a subset of $(N-1)$ players cooperate. The latter assumption did not play a role in the earlier studies, as defector types would still be better off; in the presence of punishing types, however, this may change, and the benefits forgone by T_{N-1}'s due to their highly demanding cooperation requirements may now play an important role. Finally, if we were to include errors—which we refrain from for tractability—T_{N-1}'s would always be invaded by T_{N-2}'s, for much the same reason as Tit-for-Tat players can be invaded in the two-player case.[8] For the above reasons, this paper will focus on the less demanding T_{N-2} and PT_{N-2} (i.e., T_1 and PT_1) types. In section 2.5, we demonstrate that the $(P)T_2$-types cannot invade any of the equilibria derived in the following section. To recapitulate, the agent types considered in the following model are the punishing cooperator PT_1, the conditional cooperator T_1, the reluctant cooperator DT_1, and the retaliating defector R. We will represent population states as $\mathbf{x} = (x_p, x_t, x_d, x_r)$, where indeces p, t, d, and r indicate PT_1, T_1, DT_1, and R-types, respectively.

2.3 Payoffs and equilibria

Having specified the game and the strategies to be considered, we next derive the expected payoffs corresponding to each strategy. Using table 1, it is

---

[8]See Nowak and Sigmund (1993).

Table 1: *Matching table*

i_1	i_2	i_3	$prob(i_2,i_3\|i_1)$	$u(i_1\|i_2,i_3)$	$u(i_2\|i_1,i_3)$	$u(i_3\|i_1,i_2)$
T_1	T_1	T_1	x_t^2	$\frac{b-c}{1-\pi}$		
T_1	T_1	R	$2x_tx_r$	$\frac{b/2-c}{1-\pi}$		$\frac{b}{1-\pi}$
T_1	T_1	DT_1	$2x_tx_d$	$\frac{b/2-c}{1-\pi}$		$\frac{b}{1-\pi}$
T_1	T_1	PT_1	$2x_tx_p$	$\frac{b-c}{1-\pi}$		$\frac{b-c}{1-\pi}$
T_1	R	R	x_r^2	$-c$	$\frac{b}{2}$	
T_1	R	DT_1	$2x_rx_d$	$-c$	$\frac{b}{2}$	$\frac{b}{2}$
T_1	R	PT_1	$2x_rx_p$	$\frac{b/2-c}{1-\pi}$	$b-p-k+\pi\frac{b/2}{1-\pi}$	$\frac{b}{2}-c-k-p+\pi\frac{b/2-c}{1-\pi}$
T_1	DT_1	DT_1	x_d^2	$-c$	$\frac{b}{2}$	
T_1	DT_1	PT_1	$2x_dx_p$	$\frac{b}{2}-c+\pi\frac{b-c}{1-\pi}$	$b-p+\pi\frac{b-c}{1-\pi}$	$\frac{b}{2}-c-k+\pi\frac{b-c}{1-\pi}$
T_1	PT_1	PT_1	x_p^2	$\frac{b-c}{1-\pi}$	$\frac{b-c}{1-\pi}$	
R	R	R	x_r^2	0		
R	R	DT_1	$2x_rx_d$	0		0
R	R	PT_1	$2x_rx_p$	$\frac{b}{2}-p-k$		$-c-2k-2p$
R	DT_1	DT_1	x_d^2	0	0	
R	DT_1	PT_1	$2x_dx_p$	$\frac{b}{2}-p-k+\pi\frac{b/2}{1-\pi}$	$\frac{b}{2}-p+\pi\frac{\pi b/2-c}{1-\pi^2}$	$-c-2k-p+\pi\frac{b/2-\pi c}{1-\pi^2}$
R	PT_1	PT_1	x_p^2	$b-2p-2k+\pi\frac{b}{1-\pi}$	$\frac{b}{2}-p-k+\pi\frac{b/2-c}{1-\pi}$	
DT_1	DT_1	DT_1	x_d^2	0		
DT_1	DT_1	PT_1	$2x_dx_p$	$\frac{b}{2}-p+\pi\frac{b-c}{1-\pi}$		$-c-2k+\pi\frac{b-c}{1-\pi}$
DT_1	PT_1	PT_1	x_p^2	$b-2p+\pi\frac{b-c}{1-\pi}$	$\frac{b}{2}-c-k+\pi\frac{b-c}{1-\pi}$	
PT_1	PT_1	PT_1	x_p^2	$\frac{b-c}{1-\pi}$		

Note: Probabilities with i_2 and i_3 being focal are ommitted for brevity. Payoffs for i_2 (and i_3) are only given where different from i_1 (i_2).

tedious but straightforward to determine the following payoff equations:

$$
\begin{aligned}
u(R,\mathbf{x}) &= b\left(x_t+x_p+\tfrac{\pi}{1-\pi}\left[(x_t+x_p)^2+x_px_d)\right]\right)-2(k+p)x_p \\
u(DT_1,\mathbf{x}) &= b\left(x_t+x_p+\tfrac{\pi}{1-\pi}\left[(x_t+x_p)^2+x_p(2x_d+\tfrac{\pi}{1+\pi}x_r)\right]\right)-2px_p- \\
&\quad -\tfrac{\pi c}{1-\pi}\left(2x_p-x_p^2-\tfrac{\pi}{1+\pi}2x_px_r\right) \\
u(T_1,\mathbf{x}) &= \tfrac{b}{1-\pi}(x_t+x_p+\pi x_dx_p)-\tfrac{c}{1-\pi}\left(1-\pi\left[1-x_t(2-x_t)-x_p(2-x_p)+2x_px_t\right]\right) \\
u(PT_1,\mathbf{x}) &= \tfrac{b}{1-\pi}\left(x_t+x_p+\pi\left[x_d(x_t+x_p+x_d+\tfrac{x_r}{1+\pi})\right]\right)-2kx_d-2(k+p)x_r- \\
&\quad -\tfrac{c}{1-\pi}\left(1-\pi\left[1-x_t(2-x_t)-x_p(2-x_p)+2x_px_t-x_d(x_d+2\tfrac{\pi}{1+\pi}x_r)\right]\right),
\end{aligned}
$$
$$(2)$$

where the time index τ is dropped for legibility.

Focusing on monomorphic equilibria, we immediately see that none of the four strategies selected is evolutionarily stable: R's can always be invaded by neutral drift by DT_1's and *vice versa*; the same holds true for PT_1's and T_1's, respectively. Additionally, in the absence of punisher types, T_1's can be invaded by either defecting type. At the same time, T_1's can invade any of the defecting types as well as any mix thereof, as long as the continuation probability π is sufficiently close to unity; the lower π, the higher the critical fraction of T_1's needed to invade until invasion becomes impossible altogether. The defecting types' invasion barrier against T_1's is given by

$$\hat{x}_t^{crit} = \frac{b - 2c - \sqrt{b^2 - 4c[b - c]/\pi}}{2(b - c)}, \tag{3}$$

which exists only if the continuation probability π and the public good's benefit-to-cost ratio are sufficiently large.[9]   Obviously, the ability of the conditional-cooperator type to invade a defecting population stems from the former's ability to concentrate the benefits of future cooperation predominantly on their likes, which can only pay if future interactions are sufficiently likely. On the other hand, PT_1's cannot invade R's while they can invade DT_1's, as long as their investment into establishing cooperation within an interaction is outweighed by the expected long-term benefits:

$$c + 2k < \pi \frac{b - c}{1 - \pi}. \tag{4}$$

The following proposition establishes an interesting result that is counterintuitive only at first sight.

Proposition 1. *If the continuation probability π exceeds $\pi^{max} = \frac{2p + 2k - c}{2p + 2k}$, neither cooperative strategy is stable against invasion by some defecting type.*

Proof. By (2), and as noted above,

$$u\left(T_1, (0, 1 - \varepsilon, \varepsilon, 0)\right) - u\left(DT_1, (0, 1 - \varepsilon, \varepsilon, 0)\right) = -\frac{(1 - \varepsilon^2 \pi)c - \varepsilon(1 - \varepsilon)\pi b}{1 - \pi},$$

which is negative under any value of π for small enough ε. PT_1, on the other hand, is stable against invasion by R if and only if

$$\begin{aligned} u(PT_1, (1 - \varepsilon, 0, 0, \varepsilon)) - u\left(R, (1 - \varepsilon, 0, 0, \varepsilon)\right)) = \\ 2(k + p)(1 - 2\varepsilon) - \tfrac{(1 - \varepsilon^2 \pi)c - \varepsilon(1 - \varepsilon)\pi b}{1 - \pi} > 0. \end{aligned} \tag{5}$$

Letting ε go to 0 and rearranging, we obtain the critical value from the proposition. $\square$

---

[9]For similar results, cf. Joshi (1987) and Boyd and Richerson (1988).

The intuition for why a high continuation probability π leads to punishing cooperators being invaded is analogous to why punishers can invade a population of reluctant cooperators. Either type incurs punishment costs in the first stage game to reap future benefits: the punisher by inducing cooperation and the retaliator by curtailing being sanctioned. Conversely, punishers resist invasion by retaliators if and only if conforming to the prevailing— cooperative—behaviour in all interaction stage games is less costly than incurring the one-time cost of being punished and retaliating; that is, if

$$\frac{c}{1-\pi} < 2(p+k). \tag{6}$$

Finally, DT_1's cannot invade a population of all PT_1's as long as

$$c < 2p. \tag{7}$$

This is straightforward: after the first stage game, punishers have induced DT_1's to contribute, so there are no payoff differences from the second stage game on; further, if contributing is less costly than being punished by all other players, then it pays not to try free-riding in the first stage game when all other players are punishers. Note that condition (7) may imply condition (6) if $\pi < k/(k+p)$, and *vice versa*, otherwise. This simply speaks of the fact that, when a second encounter is sufficiently unlikely, retaliating does not pay, as the probability of future gains from unpunished free-riding are close to nil. In the remainder of this article, we will concentrate on future encounters being sufficiently likely, so that (6) implies (7).

Proposition 2. *Assume $c < 2p$. Then, the subset of points in the PT_1–T_1 continuum that are stable against invasion by both defecting types is nonempty and given by $\mathbf{x}_C^{NE} = \left\{(1 - \theta(1 - \underline{x}_p), \theta(1 - \underline{x}_p), 0, 0)|\theta \in [0,1]\right\}$, where $\underline{x}_p = \max\{c/[(1-\pi)(2k+2p)], c/2p\}$, if and only if*

$$\pi \begin{cases} < \frac{2p+2k-c}{2p+2k}, & b \le \frac{2c}{\pi}, \text{ and} \\ \le \frac{2p+2k-c}{2p+2k}, & \text{otherwise.} \end{cases} \tag{8}$$

Proof. First of all, along the PT_1–T_1 continuum, all cooperating strategies obtain the same payoff of $\frac{b-c}{1-\pi}$. Equating $u(R, (x_p', 1 - x_p', 0, 0))$ as defined by (2) with $\frac{b-c}{1-\pi}$ gives $x_p' = c/[(1-\pi)(2k+2p)]$; for $u(DT_1, (x_p'', 1-x_p'', 0, 0)) = \frac{b-c}{1-\pi}$, we obtain $x_p'' = c/2p$. (8) and $c < 2p$ ensure the existence of $\mathbf{x}_C^{NE}$. By the monotonicity of defector payoffs along the PT_1–T_1 vertex, any fraction of punishers above $\underline{x}_p$ implies a lower payoff for the defecting entrants.

If $\pi > \frac{2p+2k-c}{2p+2k}$, **proposition 1** ensures $\mathbf{x}_C^{NE} = \emptyset$. Finally, if $\pi = \frac{2p+2k-c}{2p+2k}$, evolutionary stability of $\mathbf{x}_C^{NE} = \{(1,0,0,0)\}$ depends on the exact b-to-c

ratio: if $\varepsilon = 0$, $u(PT_1, \mathbf{x}_C^{NE}) - u(R, \mathbf{x}_C^{NE}) = 0$; at the same time, the left-hand side of (5) decreases in ε if and only if its derivative is negative:

$$-2c + \pi b - 2\varepsilon\pi(b - c) < 0, \tag{9}$$

obtained by substituting $\frac{c}{1-\pi} = 2(p + k)$ and rearranging. Hence, only if $b > \frac{2c}{\pi}$ is there an $\bar{\varepsilon}$ such that any ε, $0 < \varepsilon < \bar{\varepsilon}$, yields $u(PT_1, (1 - \varepsilon, 0, 0, \varepsilon)) - u(R, (1 - \varepsilon, 0, 0, \varepsilon))) > 0$. $\qquad \square$

Proposition 2 describes the set of 'fully-cooperative' equilibria and delineates the conditions for their existence. As expected, the prospects for cooperation are better the lower the net contribution costs c and the stronger the impact of a punishment action. Evidently, the 'cooperative' equilibrium component presented in **proposition 2** is not the only set of (neutrally) stable states: there is also a 'defecting component' $\mathbf{x}_D^{NE} = \left\{ (0, 0, \delta\bar{x}_d, 1 - \delta\bar{x}_d) \mid \delta \in [0, 1) \right\}$, where $\bar{x}_d$ is the highest fraction of DT_1's under which PT_1's do not obtain a positive payoff, and therefore, cannot invade the mix of retaliators and reluctant conditional cooperators.[10]

Unless the continuation probability is close to unity (and hence, T_1's could invade defectors), both components are also dynamically stable in the sense that a small perturbation away from any point but the end points of these components results in the system reverting to a point on the same component that is not too far away. The reason for this stability is that near the defecting component $\mathbf{x}_D^{NE}$, retaliation is not costly because there are so few punishers (i.e., $u(R, \mathbf{x}) \approx u(DT_1, \mathbf{x})$) while punishment is because of the large fraction of retaliators (i.e., $u(PT_1, \mathbf{x}) < u(R, \mathbf{x})$; T_1's cannot invade, anyway); near the cooperative component $\mathbf{x}_C^{NE}$, little defection happens, which means being a punisher is not costly (i.e., $u(PT_1, \mathbf{x}) \approx u(T_1, \mathbf{x})$), while defection is costly due to the high fraction of punishers (i.e., $u(R, \mathbf{x}) < u(DT_1, \mathbf{x}) < u(PT_1, \mathbf{x})$). The induced adaptations by unsuccessful invaders and the agents using the pure strategy being slightly more costly than the other pure strategy contained in the equilibrium take place at comparable time rates in the replicator dynamic. Therefore, the system moves back towards the respective equilibrium component, unless when the population approaches the respective

---

[10]This fraction is given by the solution to $\bar{x}_d = -\frac{p(1-\pi^2) + \pi(b/2 - \pi c)}{\pi^2 b - \pi c(1-\pi)} + \sqrt{\left(\frac{p(1-\pi^2) + \pi(b/2 - \pi c)}{\pi^2 b - \pi c(1-\pi)}\right)^2 + \frac{(1-\pi^2)(c+2k+2p)}{\pi^2 b - \pi c(1-\pi)}}$. This solution exists: the first term is negative because $p(1-\pi^2)$, $b/2 - \pi c$, and $\pi^2 b - \pi c(1-\pi)$ are positive: obviously, $p(1-\pi^2) > 0$; $b/2 - \pi c > 0$ because, by assumption, $(N-2)b/(N-1) > c$; and $\pi^2 b > \pi c(1-\pi)$ because, following from (4), $\pi^2 b > \pi c + 2\pi k(1-\pi)$, and, in turn, $\pi c + 2\pi k(1-\pi) > \pi c(1-\pi)$. From there, it is obvious that $\bar{x}_d > 0$; on the other hand, condition (4) also ensures $\bar{x}_d < 1$.

component's end-point.[11]

Finally, the examination of monomorphic equilibria at the outset suggests there may be another stable equilibrium component. If the conditions are such that some critical mass of T_1's can invade defecting strategies and conversely, defecting strategies can invade conditional cooperators, then there must be at least one other set of steady states $\mathbf{x}_M = \{(0, \hat{x}_t^{NE}, \rho(1 - \hat{x}_t^{NE}), (1 - \rho)(1 - \hat{x}_t^{NE}))|\rho \in [0,1]\}$, in which a fraction of T_1's coexists with a—possibly degenerate—mix of defecting types.[12] This set is defined by

$$\hat{x}_t^{NE} = \frac{b - 2c + \sqrt{b^2 - 4c[b - c]/\pi}}{2(b - c)}. \tag{10}$$

Proposition 3. *Assume $c < 2p$. If $\hat{x}_t^{NE} = \frac{b - 2c + \sqrt{b^2 - 4c[b-c]/\pi}}{2(b-c)}$ has a solution within $[0, 1]$, the subset of steady states on the T_1–DT_1–R plain that is stable against invasion by PT_1's is nonempty and given by $\mathbf{x}_3^{NE} = \{(0, \hat{x}_t^{NE}, \rho(1 - \hat{x}_t^{NE}), (1 - \rho)(1 - \hat{x}_t^{NE}))|\rho \in [0, \hat{\rho}]\}$, where $\hat{\rho} = \frac{2p + \pi(b - 2c) + \pi^2(b - 2p) - \sqrt{[2p + \pi(b - 2c) + \pi^2(b - 2p)]^2 - 8c(k + p)(1 - \pi - \pi^2 + \pi^3)}}{2c(1 - \pi)}.$*

Proof. Existence of stable states in $\mathbf{x}_3^{NE}$ follows from the fact that for $\rho = 0$, $u(PT_1, \mathbf{x}_3^{NE}) - u(T_1, \mathbf{x}_3^{NE}) = -(1 - \hat{x}_t^{NE})2(k + p) < 0$. To obtain the boundary of $\mathbf{x}_3^{NE}$, we equate $u(T_1, \mathbf{x}_3^{NE})$ to $u(PT_1, \mathbf{x}_3^{NE})$ and rearrange to obtain

$$\frac{(x_d^2 + x_d x_r)\pi b}{1 - \pi} - (x_d^2 + 2x_d x_r + x_r^2) \cdot 2k - \left(x_d^2 + \frac{2x_d x_r \pi}{1 + \pi}\right)\frac{c}{1 - \pi} - \\ -(x_d x_r + x_r^2) \cdot 2p = 0. \tag{11}$$

Recalling that $x_d = \rho(1 - \hat{x}_t^{NE})$ and $x_r = (1 - \rho)(1 - \hat{x}_t^{NE})$, (11) simplifies to

$$-\rho^2 \frac{c}{1 + \pi} + \rho\left(\frac{\pi b}{1 - \pi} - \frac{2\pi c}{1 - \pi^2} + 2p\right) - 2(k + p) = 0. \tag{12}$$

Solving (12) for ρ yields $\hat{\rho}$ from the proposition.[13] Monotonicity of PT_1- and T_1-payoffs along ρ ensures $u(PT_1, \mathbf{x}_3^{NE}) < u(T_1, \mathbf{x}_3^{NE})$ for all $\mathbf{x} \in \mathbf{x}_3^{NE}$. $\square$

---

[11]Cf. Weibull and Salomonsson, 2006, and references cited therein, for an analogous argument on very similar dynamic phenomena.

[12]If $\hat{x}_t^{crit}$ as defined in (3) is larger than 0, its combination with any mix of defecting strategies also constitutes a set of—necessarily unstable—steady states.

[13]Obviously, there may be another solution, $\hat{\rho}'$, to (12), replacing the minus in front of the square root by a plus. Yet, if $\hat{\rho}'$ exists, $\hat{\rho}' > 1$: for existence, the term under the square root must be positive, i.e., $[2p + \pi(b - 2c) + \pi^2(b - 2p)]^2 > 8c(k + p)(1 - \pi - \pi^2 + \pi^3)$. $\hat{\rho}' < 1$ obviously requires $2p + \pi(b - 2c) + \pi^2(b - 2p) < 2c(1 - \pi)$. However, $[2c(1 - \pi)]^2 < 8c(k + p)(1 - \pi - \pi^2 + \pi^3)$ under $c < 2p$ for any π. Hence, either $\hat{\rho}'$ does not exist or $\hat{\rho}' > 1$.
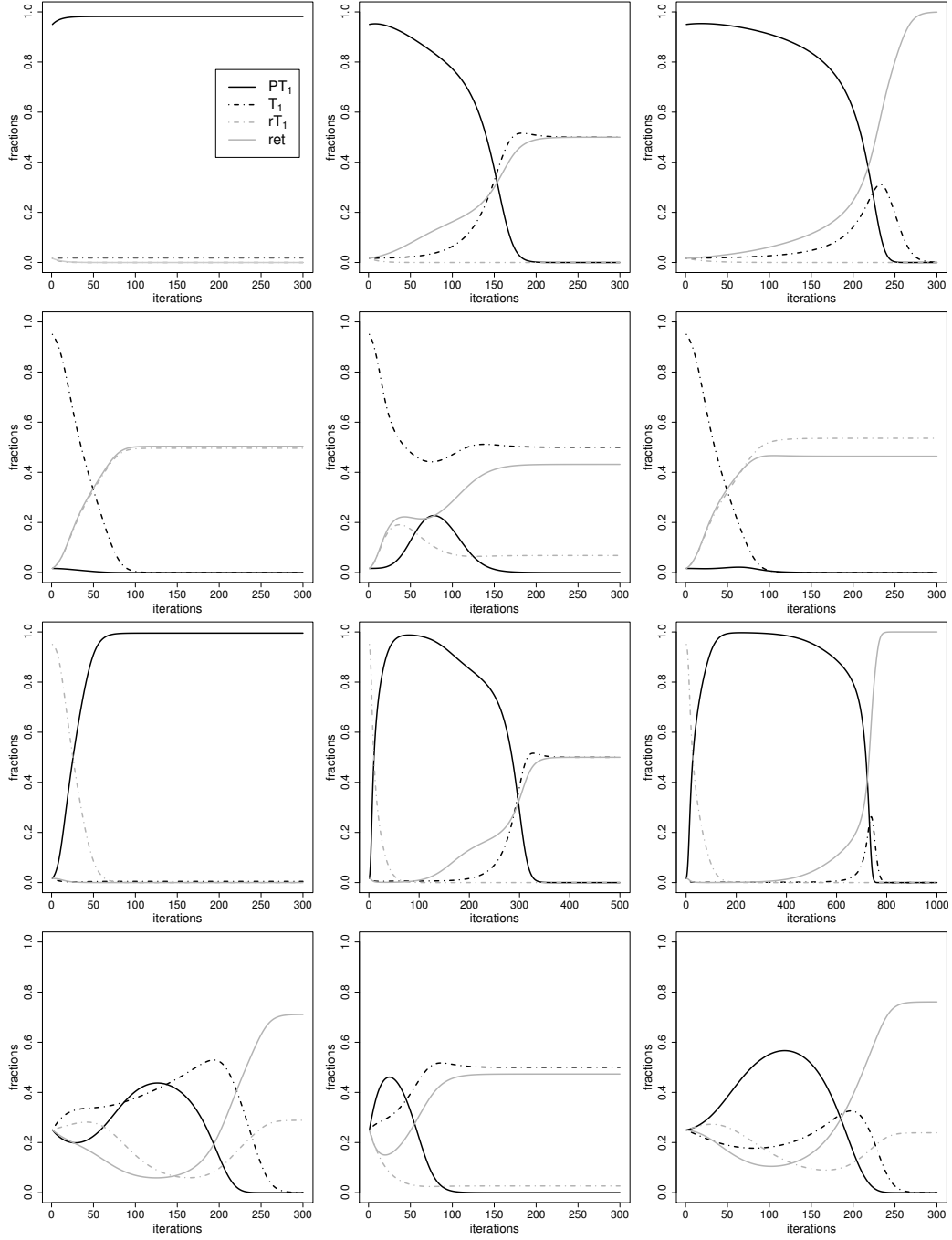
Figure 1: Evolution of types under $\pi = 3/5, k = 1/2, p = 3/2$ (first column); $\pi = 4/5, k = 1/2, p = 3/2$ (second column); $\pi = 3/5, k = 2/7, p = 6/7$ (third column); in all cases, $c = 1$, $b = 4$ (i.e., $MPCR = 2/3$). Note that in the central and right-hand third-row panels, the time-scale was adapted.

2.4 Dynamics

For a better understanding of the model dynamics, we provide exemplary model trajectories under different conditions in figure 1. As initial population states, we choose strategy distributions such that—initially equal—small shares of invading strategies compete with one incumbent strategy. We do not depict R as the incumbent strategy, as it is always stable against invasion (and would always look like the upper left-hand panel in figure 1, only with the lightest grey on top). Instead, we add the uniform mix of types as an initial state, for an illustration on the dynamics under more mixed initial states.

The first column represents the case in which all preconditions for the emergence of cooperation are favourable; in particular, the continuation probability π fulfils condition (8), and $p > c$, so that a single punisher is enough to make contributing worthwhile. The second condition is less restrictive than it may sound: punishment only has to be higher than the *net costs* of contributing, that is, contribution costs net of the benefit derived from the public good.[14] The panels in this column confirm what we have derived above: there are two types of equilibrium sets to which the population converges, a set in which cooperation prevails and one that leads to universal defection.

In the second column, we depict the case when condition (6) is violated, that is, π is high enough that it pays to be a retaliator in a population of only PT_1's. In line with **propositions 1** and **2**, a population with too high a fraction of punishers is readily invaded by retaliators. The latter may be followed by conditional cooperators if public-good benefit and continuation probability are large enough (so that the solution to (10) exists as in the depicted case). However, under some initial population states, conditional cooperators will only be able to invade a defecting population if the initial presence of PT_1's allows the fraction of T_1's to surpass the critical mass: if the initial punisher fraction is set to zero in the third central-column panel, the population is stuck in a defecting equilibrium (not depicted). In this sense, we can state:

Proposition 4. *Fix π such that $\frac{c+2k}{b+2k} < \pi < 1$. If $\mathbf{x}_3^{NE}$ is nonempty (i.e., the b-to-c ratio and π are sufficiently high), then $c \geq 2k/3$ is sufficient (but not necessary) to ensure the existence of initial states such that a punishing strategy is needed to reach the 'partially-cooperative' equilibrium component*

---

[14]For example, in a linear laboratory setting with endowments of 10 experimental-currency units and a marginal per-capita return of 0.6, this would mean that a punisher can harm any other player at least by 4 units.

$\mathbf{x}_3^{NE}$. *Notwithstanding, the punishing strategy may be driven to extinction in the ensuing equilibrium.*

Proof. Note that $\hat{x}_t^{crit}$ defined by (3) exists whenever $\mathbf{x}_3^{NE}$ is nonempty. Moreover, when PT_1's are absent, any mix of DT_1 and R is stable against invasion by T_1: by $b > c$, $\hat{x}_t^{crit} > 0$ for any $\pi < 1$. Particularly, this holds also for mixes of defector types in which DT_1 is very prevalent. Consider the special case $\mathbf{x}' = (0, 0, 1, 0)$. By (4), PT_1's can invade $\mathbf{x}'$ if $\pi > \frac{c+2k}{b+2k}$. Existence of $\mathbf{x}_3^{NE}$ obviously requires $b^2 - 4c[b-c]/\pi > 0$, which implies $\pi > \frac{c+2k}{b+2k}$ if $c > 2k/3$.[15] The final claim from the proposition is evident from the example given in the third central-column panel of figure 1. $\qquad\square$

Proposition 4 does not mean being a punisher is a dumb thing to do: the third central-column panel shows the punishing strategy may be highly successful in the initial, defecting state of the world, provided defectors are likely to be reluctant cooperators. What the extinction of punishers does mean is that it pays to switch strategies once retaliators become sufficiently common.

The right-hand column represents the case that $p < c$, the case in which the costs a single punisher can inflict on a non-contributing agent are lower than the net costs of contributing; at the same time, the continuation probability is both too high for PT_1 to be stable against invasion by retaliators and too low for an equilibrium in which conditionally cooperative and defecting types coexist. Therefore, the only attractor left is the 'defecting' equilibrium component $\mathbf{x}_D^{NE}$, in this case comprising all possible mixes of reluctant defectors and retaliators, even though punishers may initially spread and persist for some time before being driven out by retaliating players. If we reduce the continuation probability so that PT_1 is stable again, we obtain horizontally-stretched versions of the upper three first-column panels (not depicted here). However, when starting from an uniformly mixed initial population state, the population goes to pure defection straight away. This clearly indicates what should be clear from the beginning: if the expected potential benefits from future interaction are small because these interactions are unlikely to happen, and if punishers have rather limited punishment possibilities, the likelihood of a cooperative equilibrium being reached is small.

---

[15]To see this, rearrange $b^2 - 4c[b-c]/\pi > 0$ to $\pi > \frac{4cb-4c^2}{b^2}$, implying $\pi > \frac{c+2k}{b+2k}$ if and only if $\omega \equiv 3b^2c + 8bck - 4bc^2 - 8c^2k - 2b^2k > 0$. ω increases in b if $6bc + 8ck - 4c^2 - 4bk > 0$, which is fulfilled for any $b > 2c$ if $c \geq 2k/3$. Hence, $\omega > 12c^3 + 16c^2k - 8c^3 - 8c^2k - 8c^2k$ by $b > 2c$, and consequently, $\omega > 0$.

2.5   Sensitivity to strategy exclusion and mutation

In this section, we will touch briefly on two things: (i) we will illustrate that neither of the equilibria derived above would be invaded by strategies that we excluded from the analysis; and (ii), we will review what happens to the prevalence of cooperation if we allow for mutations. For the first purpose, figure 2 ist provided. It depicts trajectories for the cases when small fractions of PT_2, T_2, DT_2, and D-types are injected into the long-term equilibria (in order not to tilt the results in any direction, we chose equal fractions of DT_1's and R's in the center and right panels). What figure 2 shows is that the equilibria derived above are stable against invasion by the undominated types disregarded in our analysis. If at all, they may help PT_1's to invade a polimorphic equilibrium temporarily in the high continuation probability case (right-hand panel).
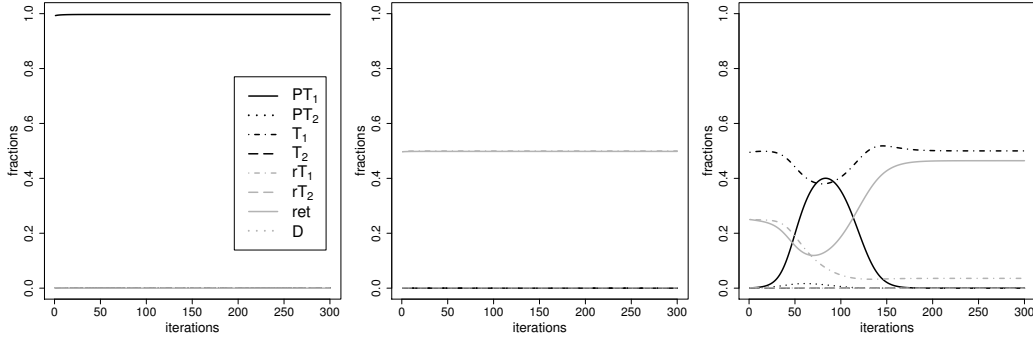

Figure 2: Stability of equilibria under $\pi = 3/5$, (left, center), and $\pi = 4/5$ (right), when formerly-excluded strategies are included; in all cases, $c = 1$, $b = 4$ (i.e., $MPCR = 2/3$), $k = 1/2$, and $p = 3/2$.

The remainder of this section is devoted to the cooperative-equilibrium cases when mutations occur. Given equations (2), it is easy to show that in a world of omnilateral defection, retaliators are almost always selected against under our assumptions when the cooperating types are maintained in small fractions by mutation.[16] The intuition is clear: the defecting types do equally well when matched with conditional cooperators; but a retaliator meeting a punishing cooperator incurs additional costs to leash back at the punisher

---

[16]The condition is given by $\frac{\pi}{1-\pi}\left[b\left(1 - 2\epsilon - \frac{1}{1+\pi}x_r\right) - c\left(2 - \varepsilon - \frac{2\pi}{1+\pi}x_r\right)\right] + 2k > 0$, where ε is the fraction of entrants per type that are produced by mutation. The left-hand side decreases in x_r when, as assumed, $b > 2c$. If we set $x_r \equiv 1 - 2\varepsilon$ and take the limit for $\varepsilon \to 0$, the left-hand condition reduces to $\pi b - 2c + \frac{2k(1-\pi^2)}{\pi} > 0$, which will be fulfilled for most relevant parameter combinations (k, π).

while substantially lowering the chances of benefitting from the public good in the future. On the other hand, with mutation keeping the defecting types alive, punishing cooperators will be selected against strongly in a world in which they are omnipresent while they would increase in numbers under the parameters chosen in this paper when everybody else is a conditional cooperator.[17] The intuition for these facts is that, when everybody is a punisher, it is costless to free-ride on punishment without losing the benefits in terms of future interactions, while when nobody else punishes, a punisher reaps higher benefits—compared to a conditional cooperator—from sporadic encounters with reluctant cooperators.

Combining the above, we would expect a circle: retaliators would be taken over by reluctant cooperators, followed suit by punishers and conditional cooperators who are, in turn, invaded by retaliators as soon as the fraction of punishers starts to fade. This is, indeed, what happens when the conditions for cooperation are favourable (i.e., when condition (8) holds) and as long as mutation is weak. We illustrate this in the center and right-hand first-row panels of figure 3. When mutation is stronger, it can induce a polimorphic equilibrium as in the upper left-hand panel of figure 3.[18]

In the second row, we depict the other case in which a positive level of co-operation was attainable, namely, when punishment was not constrained to being weak and the likelihood of future interactions is high. In the absence of mutation the population approached a polimorphic equilibrium component comprising conditional cooperators and a mix of defectors that was stable against invasion by punishers. With mutation, the equilibrium is fully mixed, as mutation counteracts the selective forces working against punishers, with the corresponding repercussions on the equilibrium levels of retaliators, conditional and reluctant cooperators. A numerical analysis suggests that the equilibrium level of punishers—and cooperation, for that matter—increases monotonically in the mutation rate in the relevant range, while the equilibrium converges smoothly towards an equilibrium in the above-mentioned component as the mutation rate approaches zero, as indicated also by the according second-row panels. Note that the effects of mutation on the cooperation level may be substantial: already for the case depicted in the central panel (i.e., for a mutation rate of one percent), first- and further-iteration cooperation rates are 60.4% and 52.9%, respectively, compared to 50% and 37.5% when no mutations occur.

---

[17]Disregarding any terms with squared mutant fractions, i.e., setting $\varepsilon^2 \equiv 0$, the comparison $u(T_1, \mathbf{x}) - u(PT_1, \mathbf{x}) > 0$ is easily transformed into $\frac{\pi b}{1-\pi} x_t - 4k - 2p < 0$, which is negative for $x_t = 0$ but positive for $x_t = 1$ under the parameter values chosen above.

[18]Note that the graphs are produced under the assumption of a very large population, so that stochastic sampling of mutants plays no role.
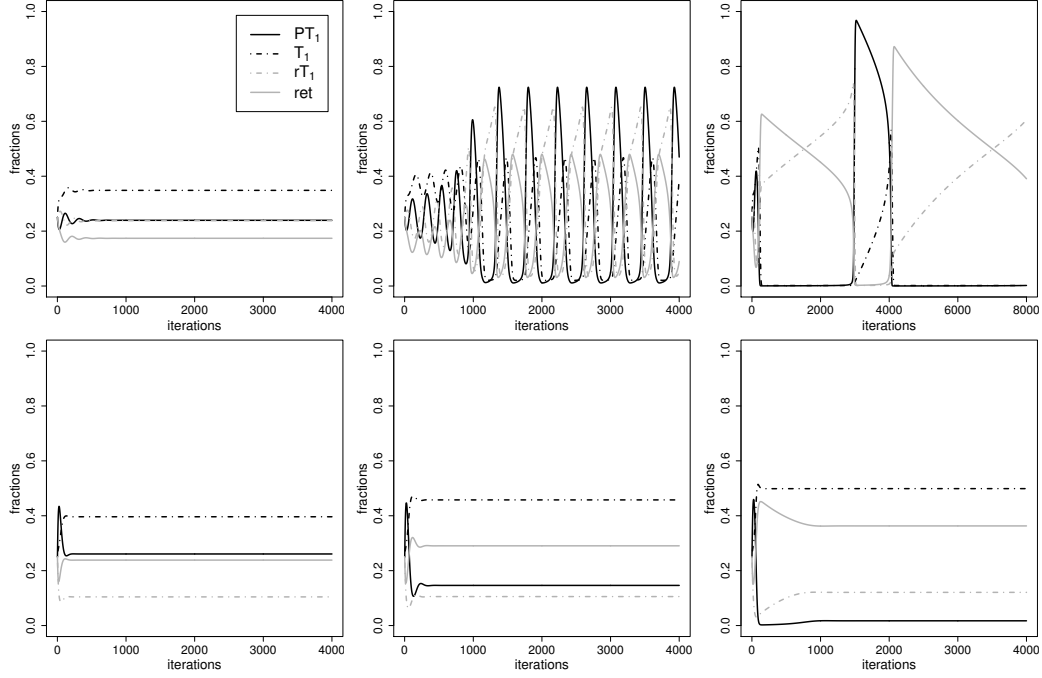
Figure 3: Dynamics for a model including mutation rates of $\mu = 0.02$ (left), $\mu = 0.01$ (center), and $\mu = 0.001$ (right), for $\pi = 3/5$ (first row) and $\pi = 4/5$ (second row); in all cases, $c = 1$, $b = 4$, $k = 1/2$, and $p = 3/2$. In the upper right-hand panel, the time scale was adapted.

Summarising, we find that, for intermediate continuation probabilities, moderate mutation puts the population on a cycle of cooperation enacted by punishment, free-riding on punishment, defection supported by retaliation, decline of retaliation, and cooperation-by-punishment again, where the periods of defection tend to be longer than those of cooperation. In turn, under high continuation probabilities as well as under high mutation rates, the population approaches a stable polimorphic equilibrium in which the mutation rate determines the cooperation level; the lower the mutation rate, the lower the level of cooperation, and the less punishing cooperators in the population. We re-state the above in the following result:

Result 1. *In the presence of moderate mutation rates, punishment plays two roles: under intermediate continuation probabilities fulfilling condition (8), it induces short periods of high cooperation levels between longer periods of omnilateral defection; when continuation probabilities are high, it is able to shift up notably the existing cooperation level in the polimorphic steady state.*

3   Summary and Discussion

Laboratory experiments suggest retaliation may be an important factor in human behaviour, and that this may have detrimental effects on cooperation. Yet, virtually all models of strong reciprocity and sanction-enforced cooperation disregard the possibility of defectors getting back at their punishers. In this paper, we provide a model that accounts for the spread of retaliation as part of a strategy within the context of a repeated social dilemma, and examine its effects on the achievable cooperation level, paying particular attention to the role of punishment in thwarting defection. In the model presented, retaliators may thrive in a society with punishers as they can discourage punishment actions while free-riding on future cooperative effort. Depending on the specific parameters, the population is predicted to end up in a cooperative state supported by punishment, in a polimorphic state in which a stable fraction of conditional cooperators coexists with defecting strategies, or a defecting state guarded against invasion by punishers through retaliation.

When mutation plays no role—or players do not experiment on their strategy even sporadically, under a different model interpretation—a punishing-cooperator strategy may be evolutionarily stable when the continuation probability is sufficiently small; when it surpasses a critical value, it pays for defecting types to invest into discouraging future punishment actions by retaliating (**proposition 1**). In this case, mirroring the earlier results by Joshi (1987) and Boyd and Richerson (1988), there will be a polimorphic equilibrium component comprised of conditional cooperators and a mix of defecting types, in which there are enough retaliating players to rule out invasion by punishing cooperators (**proposition 3**). Yet, even though in this equilibrium there are no punishers present, they may play an important role: in a world of omnilateral defection, in which sufficiently many players are reluctant cooperators, punishers may break the invasion barrior conditional cooperators face and thereby jump-start cooperation (**proposition 4**).

The limited role for punishment is accentuated when looking at the case that mutation plays a role. In this case, punishment merely induces short periods of cooperation that interrupt comparatively long periods of almost omnilateral defection when future interactions are not too likely; under a high continuation probability, it helps bring up the level of cooperation to a limited but non-negligible extent in an already partially-cooperative population (**result 1**). In this sense, punishment seems to be a facilitating device that can pave the way for and boost the power of conditional cooperation, rather than being an explanation for the evolution of cooperation of itself. A retaliation strategy, in contrast, has a place in this scenario as it conveys a benefit to its carrier-organism as soon as punishers become sufficiently wide-spread.

References

[1] Alexander, Richard D. (1979): *Darwinism and Human Affairs*. Seattle: University of Washington Press.

[2] Alexander, Richard D. (1987): *The Biology of Moral Systems*. New York: Aldine de Gruyter.

[3] Axelrod, Robert (1986): An evolutionary approach to norms. *American Political Science Review*, 80, 1095-1111.

[4] Axelrod, Robert, and William D. Hamilton (1981): The Evolution of Cooperation *Science* 211, 1390-1396.

[5] Bowles, Samuel, and Herbert Gintis (2004): The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoretical Population Biology* 65, 17-28.

[6] Boyd, Robert, and Peter J. Richerson (1985): *Culture and the Evolutionary Process*. Chicago: University of Chicago Press.

[7] Boyd, Robert, and Peter J. Richerson (1988): The evolution of reciprocity in sizable groups. *Journal of Theoretical Biology*, 132, 337-356.

[8] Boyd, Robert, and Peter J. Richerson (1992): Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology*, 13, 171-195.

[9] Carpenter, Jeffrey P. (2007): Punishing free-riders: How group size affects mutual monitoring and the provision of public goods. *Games and Economic Behavior* 60, 31-51.

[10] Carpenter, Jeffrey P., Peter H. Matthews, and Okomboli Ong'ong'a (2004): Why Punish? Social reciprocity and the enforcement of prosocial norms. *Journal of Evolutionary Economics* 14, 407-429.

[11] Cavalli-Sforza, Luigi L., and Marcus W. Feldman (1981): *Cultural Transmission and Evolution*. Princeton, NJ: Princeton University Press.

[12] Denant-Boemont, Laurent, David Masclet, and Charles Noussair (2007): Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic theory* 33, 145-167.

[13] Fehr, Ernst, and Simon Gächter (2000): Cooperation and Punishment in Public Goods Experiments. *American Economic Review* 90(4), 980-994.

REFERENCES

[14] Fehr, Ernst, and Joseph Henrich (2003): Is strong reciprocity a maladaption? On the evolutionary foundations of human altruism. In: Peter Hammerstein (ed.): *Genetic and cultural evolution of cooperation.* Cambridge, MA: MIT Press.

[15] Fowler, James H. (2005): Altruistic punishment and the origin of cooperation. *Proceedings of the National Academy of Sciences of the United States of America* 102, 7047-7049.

[16] Gintis, Herbert (2000): Strong Reciprocity and Human Sociality. *Journal of Theoretical Biology* 206, 169-179.

[17] Gintis, Herbert, Samuel Bowles, Robert Boyd, and Ernst Fehr (2003): Explaining altruistic behavior in humans. *Evolution and Human Behavior* 24, 153-172.

[18] Gintis, Herbert, Eric A. Smith, Samuel Bowles (2001): Costly Signaling and Cooperation. *Journal of Theoretical Biology* 213, 103-119.

[19] Hamilton, William D. (1964): The genetical evolution of social behaviour. I and II *Journal of Theoretical Biology* 7, 1-52.

[20] Hammerstein, Peter (2003, ed.): *Genetic and cultural evolution of cooperation.* Cambridge, MA: MIT Press.

[21] Hauert, Christoph, Arne Traulsen, Hannelore Brandt, Martin A. Nowak, and Karl Sigmund (2007): Via Freedom to Coercion: The Emergence of Costly Punishment. *Science* 316, 1905-1907.

[22] Henrich, Joseph, and Robert Boyd (2001): Why People Punish Defectors: Weak Conformist Transmission can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas. *Journal of Theoretical Biology* 208, 79-89.

[23] Janssen, Marco A., and Clint Bushman (2008): Evolution of Cooperation and Altruistic Punishment when Retaliation is Possible. *Journal of Theoretical Biology* 254, 541-545.

[24] Joshi, N.V. (1987): Evolution of cooperation by reciprocation within structured demes. *Journal of Genetics*, 66, 69-84.

[25] Nicklisch, Andreas, and Irenaeus Wolff (2011): Cooperation norms in multiple-stage punishment. *Journal of Public Economic Theory* 13, 791-827.

[26] Nikiforakis, Nikos (2008): Punishment and Counter-Punishment in Public Good Games: Can we really govern ourselves? *Journal of Public Economics* 92, 91-112.

REFERENCES

[27] Nowak, Martin A., and Karl Sigmund (1993): A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature* 364, 56-58.

[28] Nowak, Martin A., and Karl Sigmund (1998): Evolution of indirect reciprocity by image scoring. *Nature* 393, 573-577.

[29] Ohtsuki, Hisashi, Yoh Iwasa, and Martin A. Nowak (2009): Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. *Nature* 457, 79-82.

[30] Rand, David G., Joseph J. Armao IV, Mayuko Nakamaru, and Hisashi Ohtsuki (2010): Anti-social punishment can prevent the co-evolution of punishment and cooperation. *Journal of Theoretical Biology* 265, 624-632.

[31] Sethi, Rajiv (1996): Evolutionary stability and social norms. *Journal of Economic Behavior and Organization* 29, 113-140.

[32] Sober, Elliott, and David S. Wilson (1998): *Unto Others: The Evolution and Psychology of Unselfish Behavior.* Cambridge, MA: Harvard University Press.

[33] Stephens, Christopher (2005): Strong Reciprocity and the Comparative Method. *Analyse & Kritik* 27, 97-105.

[34] Sugden, Robert (1986): *The Economics of Rights, Co-operation, and Welfare.* Oxford: Blackwell Publishing Limited.

[35] Taylor, Peter D., and Leo B. Jonker (1978): Evolutionary stable strategies and game dynamics. *Mathematical Biosciences* 40, 145-156.

[36] Trivers, Robert L. (1971): The Evolution of Reciprocal Altruism. *Quarterly Review of Biology* 46, 35-57.

[37] Weibull, Jörgen W., and Marcus Salomonsson (2006): Natural selection and social preferences. *Journal of Theoretical Biology*, 239, 79-92.

[38] Wolff, Irenaeus (2009): Punishers under Fire: Retaliation and the Evolution of Cooperation. Working Paper.

[39] Yamagishi, Toshio (1986): The Provision of a Sanctioning System as a Public Good. *Journal of Personality and Social Psychology Review* 51, 110-116.

[40] Zahavi, Amotz (1975): Mate selection—A selection for a handicap. *Journal of Theoretical Biology* 53, 205-214.

Appendix: a model variant for large N

[For Online-Publication only] In the following, we provide a rough sketch of what the model would look like under large N, for an indication of whether our main results hold also for groups of more than three players. First of all, we adapt the conditionally-cooperative types to the new situation by replacing T_1, PT_1, and DT_1 by T_γ, PT_γ, and DT_γ, where $\gamma = \min\{g \in \mathbb{N} | \frac{gb}{N-1} > c\}$.[19] For simplicity, assume that the b-to-c ratio is such that $\gamma > \frac{N}{2}$. Under these assumptions, the payoff equations from (2) change to:

$$
u(R, \mathbf{x}) = E_{\mathbf{x}}\left[\frac{X_p + X_t}{N'}b - X_p(k+p)\right] + \pi\left(\sum_{j=\gamma+1}^{N-1} prob_{\mathbf{x}}(X_p + X_t = j)\left[\frac{jb}{(1-\pi)N'}+\right.\right.
$$
$$
\left.+ \sum_{k=1}^{N-1} prob_{\mathbf{x},j}(X_p = k)\frac{E_{\mathbf{x},j}(X_d)b}{(1-\pi)N'}\right] + \sum_{j=0}^{\gamma}\sum_{k=1}^{\gamma} prob_{\mathbf{x}}(X_p + X_t = j \wedge X_p = k) \cdot
$$
$$
\left.\cdot\left[\frac{E_{\mathbf{x},j,k}(X_d)b}{N'} + \sum_{l=\gamma+1}^{N-1} prob_{\mathbf{x},j}(X_d = l) \cdot E_{\mathbf{x},j,l}(1-X_r)\frac{\pi b}{(1-\pi)N'}\right]\right)
$$

$$
u(DT_\gamma, \mathbf{x}) = E_{\mathbf{x}}\left[\frac{X_p + X_t}{N'}b - X_p(p)\right] + \pi\left(\sum_{j=\gamma+1}^{N-1} prob_{\mathbf{x}}(X_p + X_t = j)\left[\frac{jb}{(1-\pi)N'}+\right.\right.
$$
$$
\left.+ \sum_{k=1}^{N-1} prob_{\mathbf{x},j}(X_p = k)\frac{E_{\mathbf{x},j}(X_d - 1)b/N' - c}{(1-\pi)}\right] + \sum_{j=0}^{\gamma}\sum_{k=1}^{\gamma} prob_{\mathbf{x}}(X_p + X_t = j \wedge X_p = k) \cdot
$$
$$
\cdot\left[\frac{E_{\mathbf{x},j,k}(X_d)b}{N'} - c + \sum_{l=\gamma+1}^{N-1}\left(prob_{\mathbf{x},j,k}(X_d = l \wedge X_r > 0) \cdot \pi\frac{E_{\mathbf{x},j,k,l}(1-X_r)b/N' - c}{(1-\pi)}\right) + \right.
$$
$$
+ prob_{\mathbf{x},j,k}(X_r = 0) \cdot \left(\mathbf{1}_{(X_p + X_d \geq \gamma)}\left[\frac{X_p b}{N'} + \frac{\pi b}{(1-\pi)}\right] + \right.
$$
$$
\left.\left.\left.+ \mathbf{1}_{(X_p + X_d < \gamma)}\frac{X_p b + \pi[(X_p + X_t)b - X_p N'p] + \pi^2(X_d b - N'c)}{(1-\pi^2)N'}\right)\right]\right)
$$

---

[19]While it would be an interesting endeavour to endogenise the fraction of cooperating others required by conditional cooperators, it goes beyond the scope of this appendix. Here, we want merely to obtain an indication for whether the main findings would carry over to a setting with larger groups.

$$u(T_\gamma, \mathbf{x}) = E_{\mathbf{x}}\left(\frac{X_p + X_t}{N'}\right)b - c + \pi\left(\sum_{j=\gamma}^{N-1} prob_{\mathbf{x}}(X_p + X_t = j)\left[\frac{jb/N' - c}{(1-\pi)} + \right.\right.$$

$$\left. + \sum_{k=1}^{N-1} prob_{\mathbf{x},j}(X_p = k)\frac{E_{\mathbf{x},j}(X_d - 1)b}{(1-\pi)N'}\right] + \sum_{j=0}^{\gamma-1}\sum_{k=1}^{\gamma-1} prob_{\mathbf{x}}(X_p + X_t = j \wedge X_p = k) \cdot$$

$$\cdot\left[\frac{E_{\mathbf{x},j,k}(X_d)b}{N'} + \sum_{l=\gamma+1}^{N-1}\left(prob_{\mathbf{x},j,k}(X_d = l \wedge X_r > 0)\cdot\frac{E_{\mathbf{x},j,k,l}(1 - X_r)b/N' - c}{(1-\pi)}\right) + \right.$$

$$+ prob_{\mathbf{x},j,k}(X_r = 0)\cdot\left(\mathbf{1}_{(X_p + X_d \geq \gamma)}\left[\frac{X_p(b/N' - p)}{N'} + \frac{\pi b}{(1-\pi)}\right] + \right.$$

$$\left.\left.\left. + \mathbf{1}_{(X_p + X_d < \gamma)}\frac{X_p(b - N'p) + \pi[(X_p + X_t)b - N'c] + \pi^2(X_d b)}{(1-\pi^2)N'}\right)\right]\right)$$

$$u(PT_\gamma, \mathbf{x}) = E_{\mathbf{x}}\left[\frac{X_p + X_t}{N'}b - c - X_d k - X_r(k+p)\right] + \pi\left(\sum_{j=\gamma}^{N-1} prob_{\mathbf{x}}(X_p + X_t = j)\cdot\right.$$

$$\cdot E_{\mathbf{x},j}\frac{(1 - X_r)b/N' - c}{1 - \pi} + \sum_{j=0}^{\gamma-1} prob_{\mathbf{x}}(X_p + X_t = j)\left[E_{\mathbf{x},j}\left(\frac{X_d b}{N'}\right) + \right.$$

$$+ \sum_{k=1}^{N-1}\sum_{l=\gamma}^{N-1} prob_{\mathbf{x},j,k,l}(X_r = k \wedge X_d = l)\cdot\pi\frac{(N-k)b/N' - c}{1 - \pi} + prob_{\mathbf{x},j}(X_r = 0)\cdot$$

$$\cdot\left(E_{\mathbf{x},j,X_r=0}\left[\frac{X_p b/N' - c}{1 - \pi} - X_t k\right] + \sum_{m=\gamma}^{N-1} prob_{\mathbf{x},j,X_r=0}(X_d + X_p = m)\cdot\frac{\pi(2N' - j - m)b}{(1-\pi)N'} + \right.$$

$$\left.\left.\left. + \sum_{m=0}^{\gamma-1} prob_{\mathbf{x},j,X_r=0}(X_d + X_p = m)\cdot\pi\frac{[N' - m + \pi(N' - j)]b/N' - [N' - j + \pi(N' - m)]k}{1 - \pi^2}\right)\right]\right),$$

$$(13)$$

where $N' = N - 1$; $prob_{\mathbf{x},i}(y)$ and $E_{\mathbf{x},i}(y)$ are short-hand for the probability and expected value, respectively, of variable y conditional on the population state x and the value of i from a higher-level probability argument (i.e., most often, outside the next set of parentheses); X_k is a random variable describing how many of the other $N - 1$ players are of type k; and

$$\mathbf{1}_{(C)} = \begin{cases} 1, & \text{if condition } C \text{ is fulfilled, and} \\ 0, & \text{otherwise.} \end{cases}$$

To get an idea of whether **propositions 1** and **2** are likely to carry over to larger groups, we assume a very large group size and abstract from stochastic sampling; the latter is warranted because both propositions refer to cases in which the fraction of first-stage cooperators is close to one, and thus, the fraction of groups with less than γ first-stage cooperators is negligible. At the same time, the abstraction from stochastic sampling induces us to make

an additional assumption: reluctant cooperators and punishers only switch to conditional cooperation if the fraction of punishers and, respectively, retaliators in the population reaches some critical fraction e, where e is of the order $1/N$. This ensures that tiny fractions of punishers (retaliators) in the population, which are unlikely to find their way into any one specific group, are not assumed to change the behaviour of reluctant cooperators (punishers) across the entire population. At the same time, assume that a lone punisher or retaliator *in a specific group* still influences the other group members' behaviour. Under these assumptions, expected payoffs can be written as:

$$
\begin{aligned}
u(R, \mathbf{x}) &= (x_p + x_t)b - x_p N'(k+p) + \pi\Bigg(\mathbf{1}_{(x_p+x_t \geq \frac{\gamma}{N'})}\left[\frac{(x_p+x_t)b}{1-\pi} + \mathbf{1}_{(x_p \geq e)}\frac{x_d b}{1-\pi}\right] + \\
&\quad + \mathbf{1}_{(x_p+x_t < \frac{\gamma}{N'})} \cdot \mathbf{1}_{(x_p \geq e)}\left[x_d b + \mathbf{1}_{(x_d \geq \frac{\gamma}{N'})} \cdot \pi\frac{(1-x_r)b}{1-\pi}\right]\Bigg)
\end{aligned}
$$

$$
\begin{aligned}
u(DT_\gamma, \mathbf{x}) &= (x_p + x_t)b - x_p N'p + \pi\Bigg(\mathbf{1}_{(x_p+x_t \geq \frac{\gamma}{N'})}\left[\frac{(x_p+x_t)b}{1-\pi} + \mathbf{1}_{(x_p \geq e)}\frac{x_d b - c}{1-\pi}\right] + \\
&\quad + \mathbf{1}_{(x_p+x_t < \frac{\gamma}{N'})} \cdot \mathbf{1}_{(x_p \geq e)}\bigg[x_d b - c + \mathbf{1}_{(x_r \geq e)} \cdot \mathbf{1}_{(x_d \geq \frac{\gamma}{N'})} \cdot \pi\frac{(1-x_r)b - c}{1-\pi} + \\
&\quad + \mathbf{1}_{(x_r < e)}\Big(\mathbf{1}_{(x_d+x_p \geq \frac{\gamma}{N'})}\left[x_p b + \pi\frac{(1-x_r)b-c}{1-\pi}\right] + \mathbf{1}_{(x_d+x_p < \frac{\gamma}{N'})} \cdot \\
&\quad \cdot \frac{x_p b + \pi[(x_p+x_t)b - x_p N'p] + \pi^2(x_d b - c)}{1-\pi^2}\Big)\bigg]\Bigg)
\end{aligned}
$$

$$
\begin{aligned}
u(T_\gamma, \mathbf{x}) &= (x_p + x_t)b - c + \pi\Bigg(\mathbf{1}_{(x_p+x_t \geq \frac{\gamma}{N'})}\left[\frac{(x_p+x_t)b - c}{1-\pi} + \mathbf{1}_{(x_p \geq e)}\frac{x_d b}{1-\pi}\right] + \\
&\quad + \mathbf{1}_{(x_p+x_t < \frac{\gamma}{N'})} \cdot \mathbf{1}_{(x_p \geq e)}\bigg[x_d b + \mathbf{1}_{(x_r \geq e)} \cdot \mathbf{1}_{(x_d \geq \frac{\gamma}{N'})} \cdot \pi\frac{(1-x_r)b - c}{1-\pi} + \\
&\quad + \mathbf{1}_{(x_r < e)}\Big(\mathbf{1}_{(x_d+x_p \geq \frac{\gamma}{N'})}\left[x_p(b - N'p) + \pi\frac{(1-x_r)b-c}{1-\pi}\right] + \\
&\quad + \mathbf{1}_{(x_d+x_p < \frac{\gamma}{N'})} \cdot \frac{x_p(b-N'p) + \pi[(x_p+x_t)b-c] + \pi^2 x_d b}{1-\pi^2}\Big)\bigg]\Bigg)
\end{aligned}
$$

$$
\begin{aligned}
u(PT_\gamma, \mathbf{x}) &= (x_p + x_t)b - c - x_d N'k - x_r N'(k+p) + \pi\Bigg(\mathbf{1}_{(x_p+x_t \geq \frac{\gamma}{N'})} \cdot \\
&\quad \cdot \left(\frac{(1-x_r)b-c}{1-\pi} - \mathbf{1}_{(x_r < e)} \cdot \frac{x_r N'(k+p)}{1-\pi}\right) + \mathbf{1}_{(x_p+x_t < \frac{\gamma}{N'})}\bigg[x_d b + \\
&\quad + \mathbf{1}_{(x_r \geq e)} \cdot \mathbf{1}_{(x_d \geq \frac{\gamma}{N'})} \cdot \pi\frac{(1-x_r)b-c}{1-\pi} + \mathbf{1}_{(x_r < e)}\Big(\frac{x_p b - c - x_r N'(k+p)}{1-\pi} - x_t N'k + \\
&\quad + \mathbf{1}_{(x_d+x_p \geq \frac{\gamma}{N'})} \cdot \pi\frac{(x_d+x_t)b}{1-\pi} + \mathbf{1}_{(x_d+x_p < \frac{\gamma}{N'})} \cdot \pi\frac{(x_t+\pi x_d)b - (x_d+\pi x_t)N'k}{1-\pi^2}\Big)\bigg]\Bigg).
\end{aligned}
$$

$$(14)$$

From (14), it is easy to compute

$$
u(T_\gamma, (0, 1-\varepsilon, \varepsilon, 0)) - u(DT_\gamma, (0, 1-\varepsilon, \varepsilon, 0)) \approx \left(x_t b - c + \pi\frac{x_t b - c}{1-\pi}\right) - \left(x_t b + \pi\frac{x_t b}{1-\pi}\right) = -\frac{c}{1-\pi}
$$

25

and

$$u(PT_\gamma, (1-\varepsilon, 0, 0, \varepsilon)) - u(R, (1-\varepsilon, 0, 0, \varepsilon)) \approx$$
$$\approx \left(x_p b - c - \varepsilon N'(k+p) + \pi \frac{x_p b - c - \varepsilon N'(k+p)}{1-\pi} \right) - \left(x_p b - (1-\varepsilon) N'(k+p) + \pi \frac{x_p b}{1-\pi} \right)$$
$$= -\frac{c}{1-\pi} + (1-2\varepsilon) N'(k+p) - \pi \frac{\varepsilon N'(k+p)}{1-\pi} \overset{\varepsilon \to \infty}{\longrightarrow} -\frac{c}{1-\pi} + N'(k+p).$$
(15)

Obviously, for $\pi > \frac{N'(k+p)-c}{N'(k+p)}$, the right-hand side of (15) is negative, and therefore, **proposition 1** from the main text will also hold for groups of more than three players: if the continuation probability is too high, neither cooperative strategy is stable against invasion by some defecting type.

Proposition 2 from the main body states there is a set of cooperative-strategy mixes with a certain minimum fraction of punishing cooperators in the mix, which is stable against invasion by both defecting types if the continuation probability does not exceed a certain value. First of all note that—as in the original model—also in a large-N variant of the model the payoff to both cooperative strategies in the relevant mixes is $\frac{b-c}{1-\pi}$. Analogously to the derivation given in the main text, we can determine the minimum fractions of punishers necessary to prevent retaliators and reluctant conditional cooperators from invading, namely $x'_p = \frac{c}{(1-\pi)N'(k+p)}$ and $x''_p = \frac{c}{N'p}$, respectively. Hence, the set of stable mixes of cooperative strategies will be non-empty also in the large-group model as long as π is small enough to ensure $x'_p \le 1$.

For the type of polimorphic equilibrium envisioned in **proposition 3**, stochastic sampling is essential. Hence, we have to resort to equations (13). Doing so, we can determine the equilibrium value of x_t for the border solution $\mathbf{x}_3^* = (0, x_t^*, 0, 1-x_t^*)$ by equating $u(T_\gamma, \mathbf{x}_3^*) = u(R, \mathbf{x}_3^*)$, and deriving

$$\frac{c}{\gamma b/N'} = \frac{prob_{\mathbf{x}_3^*}(X_t = \gamma)}{\sum_{j=\gamma}^{N'} prob_{\mathbf{x}_3^*}(X_t = \gamma)},$$

which will have a solution: the left-hand side is just below 1 by the definition of γ, and the right-hand side will converge to 1 from below as x_t^* converges to 0.[20] At the same time, punishers cannot invade, as $u(PT_\gamma, (0, x_t^*, 0, 1 - x_t^*)) - u(T_\gamma, (0, x_t^*, 0, 1 - x_t^*)) = -(1 - x_t^*)N'(k+p)$. Hence, **proposition 3**, establishing the existence of a non-empty set of polimorphic equilibria comprising conditional cooperators and certain mixtures of the defector types should also hold in the large-group model.

---

[20]Nonetheless, this exercise suggests that T_γ as defined above may not be the socially optimal T-type: the equilibrium fraction of conditional cooperators will be very low unless γ takes a higher value than assumed here.

Turning to **proposition 4**, we first of all note that for larger groups, it is extremely unlikely that conditional cooperators will be able to invade a defecting population on their own, given $u(T_\gamma, \mathbf{x}|x_p = 0, x_t = \varepsilon) - u(DT_\gamma, \mathbf{x}|x_p = 0, x_t = \varepsilon) > 0$ is equivalent to

$$\frac{c}{\gamma b/N'} < prob_\mathbf{x}(X_t = \gamma) = \pi \binom{N'}{\gamma} \varepsilon^\gamma (1 - \varepsilon)^{N'-\gamma}.$$

The left-hand side will be rather close to 1, whereas the right-hand side is likely to be near 0 for small enough ε, whatever the value of N' and $\pi, \pi \leq 1$. At the same time, focusing on the limiting case $\mathbf{x}' = (\varepsilon, 0, 1 - \varepsilon, 0)$, punishing cooperators may be able to invade a population comprised of defecting players only:

$$u(PT_\gamma, \mathbf{x}') - u(DT_\gamma, \mathbf{x}') \approx -c - (1 - \varepsilon)N'k + \varepsilon N'p + \pi \cdot prob_{\mathbf{x}'}(X_p = 0) \frac{b - c}{1 - \pi},$$

which, letting $\varepsilon \to 0$, is positive if and only if the continuation probability exceeds some critical value $\pi', \pi' = \frac{c + N'k}{b + N'k}$. Obviously, for larger group size N, the critical value will be higher, but there is no *a priori* reason to assume $\pi < \pi'$. In other words, for sufficiently high continuation probabilities π, certain population mixes of defecting types can be invaded by punishing cooperators. Moreover, under high values of π, punishing cooperators will yield to retaliators at some stage, plus there will be a polimorphic equilibrium similar to the one described in **proposition 3**. Taking the sum of these facts, it is not unlikely that also **proposition 4** will carry over to a world inhabited by groups that are larger than three.

With respect to mutations and **result 1**, finally, we note that varying the group size does not change the basic intuitions given in section 2.5. Considering that also the general characteristics of the underlying model seem to be robust to increases in N, we have no reason to expect the qualitative results pertaining to the inclusion of positive mutation rates to differ under a larger group size, either.