Counterpunishment revisited: an evolutionary approach

Wolff, Irenaeus

University of Erfurt, CEREB

June 2009

Counterpunishment revisited: an evolutionary approach[☆]

I. Wolff[a]

[a] *University of Erfurt / CEREB, Nordhäuser Straße 63, 99089 Erfurt, Germany*

Abstract

Evolutionary game theory has shown that in environments characterised by a social-dilemma situation punishment may be an adaptive behaviour. Experimental evidence closely corresponds to this finding but yields contradictory results on the cooperation-enhancing effect of punishment if players are allowed to retaliate against their punishers. The present study sets out to examine the question of whether cooperation will still be part of an evolutionary stable strategy if we allow for counterpunishment opportunities in a theoretic model and tries to reconcile the seemingly contradictory findings from the laboratory. We find that the apparent contradictions can be explained by a difference in the number of retaliation stages employed (one vs many) and even small differences in the degree of retaliativeness.

Key words: Public goods, Strong reciprocity, Conformism, Counter-punishment, Evolution of behavior
JEL: H41, C90

1. Introduction

Recent laboratory experiments have cast serious doubt on the external validity of earlier findings suggesting that punishment is a suitable solution for social-dilemma situations. Not only may there be some scepticism regarding the welfare effects of punishment, as the discussion coming to a head in Dreber et al. (2008) and finding its preliminary end in the study of Gächter et al. (2008) shows, but the very cooperation-enhancing effect seems to be 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

[☆]I would like to thank Nikos Nikiforakis, Bettina Rockenbach, and Arne Weiß for their detailed feedback that substantially helped improve an earlier version of the paper, as well as the members of CEREB for the many discussions along the way. Furthermore, I want to thank Reinhard Selten, Christian Rieck, and other participants of the 2008 conference of the German Society of Experimental Economics (GfeW) for their helpful comments and encouragement.

the cooperative outcome of earlier studies are to changes in the experimental setup. 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, incorporating the issue of counterpunishment in an evolutionary model of cooperation. More specifically, we set out to provide an answer to the question of whether strong reciprocity – with its disposition to punish non-cooperators – may still be part of the outcome of an evolutionary process when some (defecting) players engage in retaliative action. At the same time, our model illustrates a plausible explanation for the 'resurrection' of cooperation in subsequent studies of punishment and counterpunishment, such as Nikiforakis and Engelmann (2008) or Nicklisch and Wolff (2009). We show that both the number of retaliation stages and the degree of players' retaliativeness may play a critical role in determining the outcome. The task of explaining the emergence of retaliation, however, is left to future studies.

Evolutionary models of cooperation and strong reciprocity. Cooperation is a central aspect of human societies which researchers from many different fields are still paying attention to (e.g. Hauert et al., 2007, de Quervain et al., 2004, Henrich et al., 2006, or Herrmann et al., 2008). One avenue of research followed by scholars from fields as diverse as anthropology, biology, and economics is the study of evolutionary models to obtain a better understanding of how the large amount of cooperation observed in today's societies (observable e.g. in the preparation of a coordinated political response to the financial crisis of 2008) may have developed. A lot of work amongst students of the evolution of cooperation has been devoted to aspects such as kinship (Hamilton, 1964), reciprocal altruism (Trivers, 1971, Axelrod and Hamilton, 1981), costly signaling (Zahavi, 1975, Gintis et al., 2001), indirect reciprocity and reputation (Alexander, 1979 and 1987, Nowak and Sigmund, 1998), 'culture' (Cavalli-Sforza and Feldman, 1981, Boyd and Richerson, 1985), norms (Sugden, 1986, Sethi, 1996), group selection (Sober and Wilson, 1998), and strong reciprocity (Gintis, 2000, Gintis et al., 2003).[1]

This last concept of strong reciprocity has received a lot of attention, as it is unique among the mentioned theoretical solution in that it would predict cooperative behaviour in the artefactual world of an anonymous laboratory setting without repeated interaction between the same players.[2] In light of the fact that we observe a substantial degree of cooperation in these situations, the concept of

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

[2] I am indebted to Nikos Nikiforakis for the comment that other social norms may also account for this finding. While this is true, these other norms would typically require a sufficiently accurate *pre-play* recognition of the other player's type to survive in an evolutionary context (cf. Sethi, 1996). In the laboratory, however, this would seem to be a rather strong assumption.

strong reciprocity seems to be worth a closer look. The first thing to note is that the concept may have different readings, as the discussion triggered by Fehr and Henrich (2003) shows (cf. Stephens, 2005 or McKenzie Alexander, 2005). The general idea, however, is that of a behavioral 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".[3] In a prisoner's dilemma or 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 interation, unless – in the latter case – if the other player(s) is (are) in bad standing. On the other hand, strong reciprocity stipulates the punishment of defectors whenever there are punishment stages and regardless of players' anonymity.

While it has been shown that unconditional cooperation cannot easily be stabilised in such settings, 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). The intuition for this finding is rather straightforward: the payoff difference between an unconditional cooperator and a free-rider in the absence of punishers will always be positive, no matter what the type distribution in the society is. On the other hand, the payoff difference between a free-rider and a punisher may be negative if the fraction of punishers is large enough. Once free-riders are driven out of society, the payoff difference between punishers and pure cooperators goes to zero, such that punishers can survive and act as society's insurance against an invasion by free-riders. Recent contributions have gone on to show that punishing cooperators can even invade a non-cooperative society under certain conditions (Fowler, 2005, Hauert et al., 2007). This may happen if the public good is not an ideal-type public good, in the sense that people may opt out of the provision-and-benefit process completely, and doing so, earn a higher payoff than those participating in the public good under omnilateral defection. In that case, non-participants can take over a society of free-riders, being succeeded by either punishers or pure cooperators. Punishers will eventually be taken over by pure cooperators by neutral drift, who will subsequently be invaded by free-riders. Then, the cycle starts all over again. However, over time, society will spend most of the time in a cooperative state, given the transition between pure cooperators and punishers will often take a long time.

Experimental studies. Social dilemmas like those employed by the above studies are characterised by the fact that each player has a dominant strategy of deviating from the socially optimal choice in favour of the player's individual gain, and, if a punishment stage exists, of not spending any resources on punishment.

[3] 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).

While the standard game-theoretic solution of the dilemma is obvious, it is *a priori* far from clear whether this prediction corresponds to empirical observations or not, given real players' behaviour may be driven by other motivations apart from their individual material gain – such as strong reciprocity. One important source of empirical evidence to address this question is the experimental economics literature, as in comparison to the field, controlled laboratory experiments are not as prone to different interpretations of observed behaviour. Therefore, experiments suggest themselves as a suitable backdrop against which results from the evolutionary literature can be examined.[4]

As in evolutionary models, in public-good experiments without features that go beyond the simultaneous decision problem of how much to contribute to the public good (translating into 'all' or 'nothing' in the prisoner's-dilemma case) and without information on the other players, cooperation cannot be sustained over long rounds.[5] In contrast, Yamagishi (1986) and Fehr and Gächter (2000, 2002) were able to show that punishment was very effective in inducing cooperation in public-good games.[6] This result has been replicated e.g. by Nikiforakis (2008) and Denant-Boemont et al. (2007; henceforth, DNM). However, both latter studies go on to show that this finding is valid only when punishment is restricted to a single punishment stage following the contribution stage. If (only) punished players can hit back, counterpunishment seems to eliminate the contribution-enhancing effect of single-stage punishment. DNM go on to show that for a second punishment stage without restrictions on punishment behaviour, cooperation is still reduced when compared to the single-stage punishment case, albeit not as much as in the counterpunishment-only situation. Adding further punishment stages leads to contradicting results: while in the study by Denant-Boemont et al., the level of cooperation is as low as in the treatment with two punishment stages, Nikiforakis and Engelmann (2008, henceforth NE) observe a cooperation level that is as high as in the traditional single-punishment-stage design. Nicklisch and Wolff (2009, NW), focusing on a different question, only run a multiple-stage-punishment treatment, finding stable, non-decreasing average contributions at around half the endowment.

Previous studies and our model. Three important differences in the experimental setup come to mind that may account for the different findings:

(i) the studies employed different punishment technologies: DNM and NW use the technology employed in Fehr and Gächter (2000), while NE make use of a linear $1:2$ punishment technology similar to the linear $1:3$ technology in Fehr and Gächter (2002),[7]

[4] This approach is not new: for other examples, cf. e.g. the studies by Gintis et al. (2003), Carpenter et al. (2004), or Carpenter (2007).

[5] Cf. the extensive survey in Ledyard (1995), or the summary of the "core facts" in Ostrom (2000, p. 140).

[6] Cf. Ostrom et al. (1992) for a similar result in a commons dilemma.

[7] As Nikos Nikiforakis pointed out to me, it remains an unresolved question whether an

(ii) the end of each stage game is determined in different ways: there is an exogenously-fixed number of 5 punishment stages in the case of DNM, and an endogenous number of stages in NE and NW;[8] and

(iii) the subject pools are different: DNM conducted their experiment in Rennes (France), NE in London (UK), and NW in Bonn (Germany).

In this paper, we will focus on explanation (iii), even though we cannot rule out any of the above differences as a possible explanation for the diverging findings. However, (iii) seems especially interesting in light of the results of Herrmann et al. (2008), who find that the inclination to punish anti- as well as pro-social actions can differ considerably across societies. From our model, we will find that such population differences need not be large for substantial changes in the likelihood of a cooperative outcome.

On a more general level, another important question arises in light of the contradictory laboratory findings: can the high degree of cooperation among humans still be explained by the possibility to punish defecting society members, as the results of Nikiforakis and Engelmann suggest, or is this a rather unlikely explanation, as the earlier findings of Denant-Boemont et al. seem to show? Bearing in mind that the standard approach in the literature on the evolution of cooperation holds that "while free riders occasionally punish cooperators, they do so rarely enough that we [can] restrict the ability to punishment to cooperators,"[9] the answer to this question is far from clear. If at all, models of cooperation have included additional punishment stages reserved to sanction enforcement (Henrich and Boyd, 2001). However, the experimental counterpunishment studies show that retaliation is a force to factor in, whereas evidence for second-order punishment is rather weak (e.g., Nicklisch and Wolff, 2009). Hence, the question to be answered is whether we can really abstract from retaliation when explaining cooperation by punishment. More specifically, would strong reciprocity with its disposition to punish non-cooperators still be part of an evolutionary stable outcome when some (defecting) players engage in counterpunishment?

Henrich and Henrich (2006) identify the main reasons for the stabilisation of cooperation through punishment to work in the classical evolutionary models:

(I) "punishers don't have to pay the costs of punishing very often if being punished is more costly than the costs associated with sticking to the norm – (...) punishers need only punish occasional deviants;

(II) the cost of punishing is small (and probably ambiguous), so conformist

increase in the punishment-efficiency parameter from $1:2$ to $1:3$ would lead to an increasing trend in average contributions for multiple punishment stages as it does in the case without counterpunishment, cf. Nikiforakis and Normann (2008). While such a shift could make enforcement more effective, it would also render backlashing defectors a more dire threat.

[8] An important consequence of this difference is that, in DNM, punishment could be delayed strategically until the final punishment stage to forego retaliation.

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

transmission can overcome it, and keep a strategy of punishing stable in the social group;[10] and

(III) when punishing norm violators is common, everyone tends to adhere to the norm because the costs of being punished for violating the norm exceed the costs of sticking with the norm."

While introducing retaliation opportunities will leave (I) and (III) unaffected, the cost of punishing will increase when some of the punished strike back.[11] Hence, one would conjecture the disadvantage to be overcome by conformist transmission or intergroup conflict will increase. Indeed, we find the retaliation frequency determines the set of initial type-distributions attaining the cooperative outcome. However, we show that the introduction of retaliation possibilities may act in either direction: for high frequencies of counterpunishment, the basin of attraction of the punishment fixed point decreases as expected, while it is *expanded* for low such frequencies. This is due to the attenuation of actual payoff differences by an aggravation of the worst-case scenario, against which any payoff differences are evaluated.

For the purpose of our paper, we shortly introduce a baseline model without counterpunishment opportunities as our benchmark in section 2 and discuss its main properties before we proceed to introduce retaliation in section 3. In this section, we will contrast the results of our model for low respectively high counterpunishment frequencies to the benchmark case and subsequently address the question of whether a cooperative society and positive retaliativeness levels are mutually exclusive. In section 4, we discuss our findings relating them to the experimental findings that provided an essential part of the motivation for undertaking the present study, as well as to the results of existing evolutionary models of cooperation. We conclude in section 5.

2. The baseline model

In our basic assumptions, we will closely follow the model presented by Henrich and Boyd (2001) and modified in Boyd et al. (2003). Assume the life of an individual predominantly consists of decisions on cooperative behaviour and there are no external gains from signalling (for example through ancillary games, as in Gintis et al., 2001). While we are well-aware of the importance of partner choice possibilities (Hruschka and Henrich, 2006) as well as reputation (for a review, cf. Nowak and Sigmund, 2005), we abstract from both for the purpose of this paper, given both aspects are ruled out in the experimental settings we contrast our findings against.

At each moment in time, groups consisting of N individuals are randomly drawn from a very large population. Interaction takes place in the form of a one-

[10] They further point out that a conformist bias is not necessary when there is intergroup conflict as Boyd et al. (2003) have shown.

[11] For empirical support of this argument, cf. Nikiforakis (2008) or Nikiforakis and Engelmann (2008).

shot game consisting of two stages. In stage 0, the contribution stage, agents play a symmetric N-person binary public-good game. A cooperating player incurs a cost of c to convey a benefit of b/N to every member of the group, where $b/N < c < b$, 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/N and inducing a damage of p/N for the player punished. We assume there are three types of agents: (i) defectors who do not contribute to the public good, (ii) cooperators who contribute c unless they make a mistake with an error probability e, and (iii) punishers (strong reciprocators) who behave as cooperators in stage 0 and punish non-contributors unless they make a mistake with the same error probability e. This assymmetric treatment regarding the proneness to errors was introduced by Henrich and Boyd (2001), the intuition being that players do not erroneously exert an effort (given they have already chosen not to) but may well fail to do so, be it for external reasons or forgetfulness. While from our point of view it is plausible to assume defectors do not cooperate by chance, non-punishers could well be expected to erroneously punish with a certain probability – think of an untargeted, casual remark that is perceived as a chastisement for a non-cooperative behaviour the speaker might not even have been aware of. We abstract from this for simplicity, and for better comparability with Henrich and Boyd's results. We also abstract from errors of perception, for the same reasons.

Denoting the fraction of defectors, cooperators and punishers by π_D, π_C, and π_P, respectively, and omitting the benefit-term for ease of exposition, we have the following expected payoffs B_i:[12]

$$\begin{array}{rcl} B_D &=& -(1-e)\pi_P p, \\ B_C &=& -(1-e)c - e(1-e)\pi_P p = -(1-e)c + eB_D \\ B_P &=& -(1-e)c - e(1-e)\pi_P p - (1-e)[1-(\pi_P+\pi_C)(1-e)]k. \\ &=& B_C - (1-e)[1-(\pi_P+\pi_C)(1-e)]k \end{array} \tag{1}$$

Having established the basic game, we now proceed to derive the replicator dynamics for this case.

2.1. Replicator dynamics of the baseline game

To derive the dynamics, we assume that after playing the game, agents reproduce asexually and die. Their offspring first takes on the parent's type, but additionally gets to know a random type's average payoff and frequency of occurrence from the recent round. Alternatively, we may interpret the situation as agents running into another agent every certain time period and reconsidering their type. With probabilites α and $(1-\alpha)$ an agent chooses a learning rule, comparing either the frequency of the other agent's type, π_j, to their own, π_i, or

[12] We can omit the public-good benefit in the payoff equations without loss of generality because this term is the same for all players and for a type's fitness (or attractiveness, in our interpretation of the model), only payoff *differences* are relevant.

expected payoffs B_j versus B_i. Then, we take the probability of a type-switch to be equal to

$$prob(j|i,j) = \frac{1}{2}\big[1 + (1-\alpha)\beta(B_j - B_i) + \alpha(\pi_j - \pi_i)\big], \tag{2}$$

where β is the inverted largest-possible payoff difference between player types and thus normalises payoff differences to lie within the intervall $[-1;1]$, while $prob(j|i,j)$ is the probability that an agent of type i turns into a j-type given they meet.[13]

While it is not our purpose to enter into an extensive discussion on whether a conformist learning rule like the one our agents perform in $\alpha * 100$ percent of the cases is a suitable description of human behavior, a few words seem warranted.[14] The question comes into mind of why an agent capable of calculating expected payoffs as well as accumulating enough information to have a perfect knowledge of the society's type distribution would follow a rule that is as simple as 'do what the others do'. There are four possible answers: (i) the agent may have a preference for not standing out, for not appearing deviant,[15] or even derive utility from "simply making the same choice as one's reference group";[16] (ii) while the agent may be *capable* of accumulating the necessary information, she may face information costs that may make it a good choice to avoid gathering the information and to rely on others' example;[17] (iii) if we assume the agents do not always obtain all the information needed to calculate expected payoffs, or are not capable of doing so generally receiving information on the payoffs from some sort of agency, $(1-\alpha)$ can be interpreted as the probability with which the agent has this information. In case the agent does not, it seems plausible to use frequencies (which may be easier to obtain) as the best information available on a strategy's fitness; finally, (iv) Andrés Guzmán et al. (2007) show that a conformist learning rule may not only enhance cooperation but increase the fitness of groups and thus be adaptive in environments characterised by cooperative dilemmas and (however rare) intergroup conflict.[18] As our focus does not lie on the question of which learning-rules are going to be employed

[13] Equation (2) assumes that, if neither payoffs nor frequencies differ between the types, the individual will acquire either type with probability 0.5. In other words, there is no *status-quo* bias towards his inherited (or past) type. We admit that this will be a strong assumption in many contexts; we nevertheless follow Henrich and Boyd (2001) in this respect, as it renders the model easier to handle and more traceable. We conjecture that a deviation from this assumption would not change the qualitative results but merely slow down the model dynamics.

[14] For a more thorough discussion of this issue, cf. e.g. Henrich and Henrich (2006).

[15] Ibd.; also, cf. Zafar (2009).

[16] Zafar (2009, p. 1); in an public-good experiment with public-account payoffs going to a well-known charity, he finds that conformism plays an important role in subject behaviour.

[17] Cf. e.g. Henrich and McElreath (2003), or Richerson and Boyd (2005).

[18] Note that in their case agents do not mix learning rules – they are either payoff-oriented *or* conformist learners. In our case, this would lead to six different types in the basic model and twelve in the retaliation model, increasing the model complexity beyond reasonable limits.

by the agents, we abstract from this issue and rely on the results of the earlier studies mentioned, simply assuming agents are at least to a small extend driven by a conformist bias.

With the given average payoffs and switching probabilities we can now proceed to calculate the expected period-to-period change in type frequencies for any given type distribution.[19] Doing so, we obtain the following difference equations describing the expected fraction change:

$$\Delta_i = \pi_i(1-\pi_i)\Big((1-\alpha)\beta\big(B_i - \Sigma_{j\neq i}\tfrac{\pi_j B_j}{1-\pi_i}\big) + \alpha\big[\pi_i - \Sigma_{j\neq i}\tfrac{\pi_j^2}{1-\pi_i}\big]\Big), \qquad (3)$$

for all $i, j \in \{P, C, D\}$. Note that equation (3) holds for any number of strategies: for two strategies, it boils down to the replicator dynamics used by Henrich and Boyd (2001), while we will make use of the three- and six-strategy versions for our baseline and retaliation models, respectively. Furthermore, in contrast to the standard replicator dynamics as introduced by Taylor and Jonker (1978), for the evolutionary-stability analysis it is generally not irrelevant whether we think of a mix of invading types or the corresponding mixed-strategy invader. For our replicator dynamics this only holds for π_i sufficiently close to one: in that case we can set $\pi_i - \Sigma_{j\neq i}\frac{\pi_j^2}{1-\pi_i} \equiv 1$. To ensure evolutionary stability of a pure strategy, it is then sufficient to require that $\alpha > -(1-\alpha)\beta(B_i - \Sigma_{j\neq i}\frac{\pi_j B_j}{1-\pi_i})$ even for $\boldsymbol{\pi'_{-i}} = arg\ max_{\boldsymbol{\pi_{-i}}}\big(-B_i + \Sigma_{j\neq i}\frac{\pi_j}{1-\pi_i}B_j\big)$ subject to $\Sigma_{j\neq i}\pi_j = \varepsilon \approx 0$, which is equivalent to requiring the conformist bias to be strong enough to outweigh the normalised payoff disadvantage $i-$type players face when paired with an optimally mixing player.

Out of the three trivial fixed points in the baseline model, $P = (\pi_P = 1, \pi_C = 0, \pi_D = 0)$, $C = (0, 1, 0)$, and $D = (0, 0, 1)$, D is evolutionary stable (ES, i.e. defection is an evolutionary stable strategy, or ESS) always, while P is ES only for αs that exceed a minimum of

$$\hat{\alpha}^{base} = \begin{cases} \frac{ek}{c+k(1+e)}, & c < (1-e)p \\ \frac{c-(1-e)p+ek}{2c-(1-e)p+k(1+e)}, & \text{otherwise.} \end{cases} \qquad (4)$$

Before we go on examining the remaining fixed points in terms of evolutionary stability, let us look at the implications of equation (4) a little closer. The two cases distinguish two kinds of situations, one in which contribution enforcement is possible, $c < (1-e)p$, and one in which it is not. Recall that, for $c < (1-e)p$ to hold, every single punisher only has to inflect a damage of $(1-e)p/N$.[20] This

[19] In taking expected values, we abstract from stochastic influences, once again for reasons of comparability and simplicity. We also stick to the discrete-time modelling for similar reasons, and because it is closer to the experimental data.

[20] From this argument, we see that comparing different group sizes N does not have an influence on our results only if we assume punishers adjust their per-defector punishment expenditures $\kappa(N) = k/N$. What $c > (1-e)p$ describes, then, is the case when the number of

is a viable option in the vast majority of studies on punishment in public-good settings, notable exceptions being Decker et al. (2003) and treatments in Egas and Riedl (2008) and Nikiforakis and Normann (2008), who explicitly address the question of a very inefficient punishment technology.[21]

Proceeding with our stability analysis, we find that C will be ES only for unrealistically high αs ($\hat{\alpha} = c/(2c + k) \gg 1/3$). Apart from the trivial fixed points P, C, and D, there are two mixes of two strategies each (P and D, and P and C, respectively), plus, for certain values of alpha, an additional fully mixed fixed point.[22] None of them is stable, however, which follows directly from the stability of the trivial solutions.

In summary, we replicate the main result from Henrich and Boyd (2001):[23]

Result 1. While pure cooperation is not evolutionary stable under realistic values of a conformist bias, a cooperative outcome supported by punishment may well be.

Due to the complex nature of our dynamic system which unfortunately does not yield the readily solvable linearities of the model presented by Henrich and Boyd, in the following we resort to a numerical analysis of our model. Figure (1) shows the minimum fraction of punishers for the frequency of punishers not to decrease under different values of α for both a) low and b) high contribution costs.[24] While it is not our primary goal to provide a model tailored to give a detailed explaination for experimental-subject behaviour, we do want to compare their qualitative results with ours. For this reason, we chose our parameter values as close as possible to those often used in experiments for better comparability of results: the punishment technology is linear with a k-to-p ratio of $1 : 3$;[25] players interact in groups of $N = 4$, making mistakes in one percent of the cases.[26]

---

players is not large enough to make the expenditure $\hat{\kappa}$ s.t. $c = (1 - e)\tau\hat{\kappa}N$ affordable, where τ is the $(1 : \tau)$ punishment technology's efficiency parameter.

[21] I am grateful to Nikos Nikiforakis for pointing me to the latter two studies.

[22] Unfortunately, this equilibrium cannot easily be derived analytically, as the corresponding quadratic equations have no solution in $\mathbb{R}$ under most parameter combinations, apart from being too complex for any meaningful interpretation. We therefore had to resort to a numerical analysis to obtain this result.

[23] Henrich and Boyd obtain the result for even smaller conformist biases, as their result rests on an n^{th}-order punishment argument.

[24] Note that these are not the basins of attraction: in the blue case for low contribution costs ($c = 2$), for example, the basin of attraction of P is but an ε-environment of P.

[25] This technology has also been employed by evolutionary studies, such as Hauert et al. (2007). Other studies like Carpenter (2007) or Dreber et al. (2008) provide their agents with more powerful punishment technologies. For the purpose of our analysis, however, we stick to the widely used 1:3 technology, as this gives us a more conservative estimate of whether and under what conditions P can be stabilised.

[26] Increasing the error parameter to $e = 0.1$ does not change the qualitative results. The same holds true for changing the number of players within a group. What may result is a shift between the situation in which contributions are enforceable and when they are not, cf. footnote 20. The effects of a deterioration of monitoring possibilities that may go hand in hand with an increase in group sizes are studied in Carpenter (2007).

Under the given parameters for P to be evolutionary stable, the conformist bias need not be larger than $\hat{\alpha}^{base} \approx 0.0033$, for contribution costs such that being a punisher means investing an equivalent of one sixth of the contribution costs in a non-contributor's punishment (the left case in figures 1 and 2). On the other hand, for contribution costs such that the ratio of per-defector punishment costs and contribution costs is one tenth (corresponding to the right hand sides in figures 1 and 2), the conformist bias needs to be as large as $\hat{\alpha}^{base} \approx 0.054$. Even though a fraction of conformist learning of 5.4 percent does not seem unthinkable, these numbers illustrate an important point: the fraction of conformist learning necessary to stabilise punishment is highly sensitive to changes in the costs of contribution. More specifically, from equation (4) we obtain $\frac{\delta\hat{\alpha}}{\delta c} < 0$ for $c < (1-e)p$, given P has to be stabilised against invading cooperators, whereas $\frac{\delta\hat{\alpha}}{\delta c} > 0$ for $c > (1-e)p$, as in this case, it is the defectors who are most likely to invade a punishing society. However, the sensitivity of $\hat{\alpha}$ to changes in c is not symmetric: while the critical conformist bias hardly changes when enforcement is possible, $c < (1-e)p$, it increases more than proportionally with contribution costs for a wide range of parameters under $c > (1-e)p$.

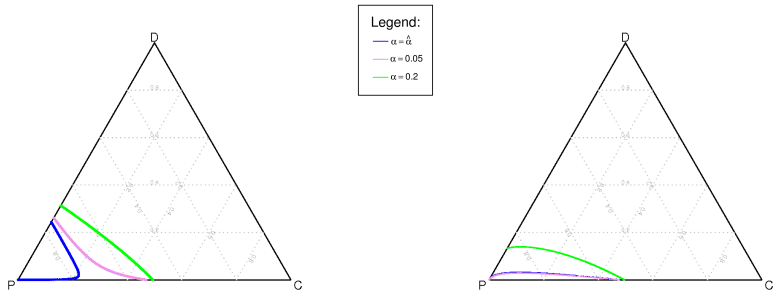

Figure 1: Minimum fractions of punishers such that $\Delta_P \geq 0$ for given values of α; a) for a low c-to-p ratio (left), and b) for a high such ratio (right).

The mixture plots in figure 1 further illustrate two points: (i) a higher conformist bias makes it easier for punishers to survive as a group (which we already know from our derivation of $\hat{\alpha}^{base}$), and (ii) for high contribution costs c, a punishing society will be invaded by defectors directly, whereas for a modest c (i.e. a lower c-to-p ratio), it will be invaded by cooperators first and only then be taken over by defectors: on the low-c graph, a small step away from P on the $\overline{PC}$-line carries society away from the punishing fixed point while defectors are still absent, whereas a small step on $\overline{PD}$ is reverted; and *vice versa* for the high-c case. As discussed above, this is reflected in the equation determining the minimum-required conformist bias for P to be ES (equation (4)). Figure 2 shows the dynamics for a fraction of conformist learning of $\alpha = 0.2$.

Figure 2 illustrates a significant difference between situations in which cooperation can be enforced by punishment and those when it cannot: for lower contribution costs c, not only the areas of positive π_P changes (cf. figure 1)

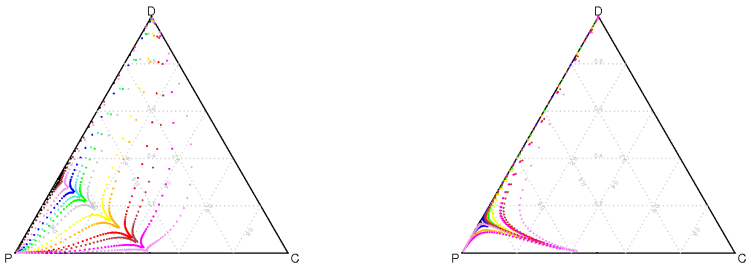

Figure 2: Dynamics for $\alpha = 0.2$ for $c = 2 < 0.99 \cdot 3 = (1-e)p$ (left) and $c = \frac{10}{3} > 0.99 \cdot 3 = (1-e)p$ (right).

but also the basin of attraction of the fixed point P is substantially larger than for higher such costs. What follows is that it is far more likely to end up in a cooperating steady state if contribution costs are low relative to the effect of punishment, another finding that is in line with what we would expect, given the payoff differences between defectors and the other types emanating from the first two stages are smaller in this case.

In the next section, we will present our extension to the baseline game introducing retaliative stages. We then proceed to discuss how the dynamics change for different initial levels of retaliativeness and what this may tell us about the likelihood of different outcomes in varying environments.

3. The retaliation game

The retaliation game is like the baseline game with n additional punishment stages. Behaviour on these stages is guided by a 'gene' (in our alternative interpretation, a behavioural rule) that controls players' retaliativeness. Given the gene may have different consequences for different player types, paying attention to it means increasing the number of player types to six - a retaliative and an "accommodating", tolerant variant of each player type introduced in the baseline model. Let $\pi_{i,r}$ ($\pi_{i,a}$) denote the fraction of i-type players with(out) a disposition to hit back in case of being punished, while π_i will still denote the total fraction of that type, i.e. $\pi_i = \pi_{i,r} + \pi_{i,a}$. Applying the same logic as before, a retaliator may make a mistake in exerting punitive effort with probability e. Evolution will proceed according to the same principles as the types in the baseline model do.

Let us first analyse how the additional stages affect the payoffs of the accommodating types. As these types do not get caught up in battles of punishment and counterpunishment, it is obvious that the expected payoff of accommodating defectors and cooperators, $B_{D,a}$ and $B_{C,a}$, is equal to that of the corresponding baseline-model types, B_D and B_C from equation (1), respectively. Only the

payoff of tolerant punishers will undergo a slight change, as they will now face counterpunishment actions by punished retaliators:

$$\begin{aligned} B_{D,a} &= -(1-e)\pi_P p = B_D \\ B_{C,a} &= -(1-e)c + eB_{D,a} = B_C \\ B_{P,a} &= B_{C,a} - (1-e)\Big([1-(\pi_C+\pi_P)(1-e)]k + \\ &\qquad +[e(\pi_{C,r}+\pi_{P,r})+\pi_{D,r}](1-e)p\Big) \\ &= B_P - (1-e)[e(\pi_{C,r}+\pi_{P,r})+\pi_{D,r}](1-e)p\,. \end{aligned}$$

On the other hand, retaliative behaviour gives rise to costs that depend on the number of additional punishment stages n. In our evolutionary-stability analysis, we will contrast two scenarios: the baseline model ($n = 0$) and the general case in which we have $n = R$ additional punishment stages. While the former scenario corresponds to the setup of Fehr and Gächter (2002) in the experimental literature, the latter can be used to represent a Nikiforakis-(2008)-like setup for $n = 1$, or the *6SFI* treatment in Denant-Boemont et al. (2007), for $n = 4$. Larger numbers of n may be thought of as corresponding to the case of an endogenous number of retaliation stages as in Nicklisch and Wolff (2009) or Nikiforakis and Engelmann (2008).[27] The *additional* payoffs from $n = R$, $R \geq 2$, retaliation stages are given by

$$\begin{aligned} \hat{B}_{D,r} &= -(1-e)^2\pi_{P,a}k - (1-e)^2\pi_{P,r}\Big(\Sigma_{r=1}^{R}[1-\mathbf{1}_E(r)](1-e)^{r-1}k + \\ &\qquad + \Sigma_{s=1}^{R}\,\mathbf{1}_E(s)(1-e)^{s-1}p\Big) \\ \hat{B}_{C,r} &= e\hat{B}_{D,r} \\ \hat{B}_{P,r} &= e\hat{B}_{D,r} - (1-e)^2\big[e(\pi_{C,r}+\pi_{P,r})+\pi_{D,r}\big]\cdot \\ &\qquad \cdot\Big(\Sigma_{r=1}^{R}\,\mathbf{1}_E(r)(1-e)^{r-1}k + \Sigma_{s=2}^{R}[1-\mathbf{1}_E(s)](1-e)^{s-1}p\Big), \end{aligned}$$

where

$$\mathbf{1}_E(x) = \begin{cases} 1, & (x \bmod 2) = 0, \\ \\ 0, & \text{otherwise.} \end{cases}$$

The additional payoff of a defector from retaliation, $\hat{B}_{D,r}$, is the expected cost of punishing back a non-retaliating punisher, $-(1-e)^2\pi_{P,a}k$, plus the expected cost of alternating counterpunishment against and being retaliated against by a retaliating punisher, where the probability of later stages diminishes due to the error probability e. A retaliating cooperator and a retaliating punisher face

[27] The largest number of *endogenous* punishment stages observed in the study of Nicklisch and Wolff (2009) was eight, followed by one observation with six. These eight (six) punishment stages boil down to one punishment plus three (two) retaliation stages if we discount rounds due to "follow-up" punishment actions directed at the same player as a punishment action on the directly preceding stage. Note that in the study discussed, and unlike in Denant-Boemont et al. (2007), this cannot be explained by subjects trying to save on costs by spreading their punishment actions over various punishment stages.

the same additional costs as the retaliating defector whenever they mistakenly defect, and the punisher faces additional costs from counterpunishing those who retaliated against his or her first-order punishment, as well as the ensuing feud. Together with the corresponding replicator dynamics equation (3) which we already know from our treatment of the baseline model, we now proceed to analyse the changes that result for our dynamic system as a whole under the different scenarios.

Each baseline-model fixed point corresponds to two different fixed points in the retaliation model, one for the retaliating and one for the accommodating variant of each type. Bearing in mind the general question we set out to answer, that of whether a cooperative outcome can be stabilised in the presence of retaliation opportunities, we have four candidate fixed points to consider:[28]

1. *accommodating punishers.* First, it is easy to see that for the accommodating variant of the punisher type to be an ESS, the requirements on the conformist bias α will be less strong than for retaliating punishment, given $B_{P,r} \leq B_{P,a} + e\hat{B}_{D,r}$ and $\hat{B}_{D,r}|_{\pi_P>0} < 0$. Furthermore, an analogous argument shows there will not be any retaliation by the optimally mixing invasion candidate. But then, we are faced with the same situation as in the baseline model. Hence, we directly obtain

$$\hat{\alpha}_{P,a} = \begin{cases} \frac{ek}{\hat{B}+ek}, & c < (1-e)p \\ \\ \frac{c-(1-e)p+ek}{\hat{B}+c-(1-e)p+ek}, & \text{otherwise,} \end{cases} \tag{5}$$

where

$$\hat{B} = \begin{cases} -\frac{1}{1-e}B_{P,r}|_{\pi_{D,r}=1}, & 2c+k > \\ & \quad > (1-e)^R\Big(\mathbf{1}_E(R)p + \\ & \quad\quad +[1-\mathbf{1}_E(R)]k\Big) + E \\ -\frac{1}{1-e}(B_{D,r}-B_{C,a})|_{\pi_{P,r}=1}, & \text{otherwise,} \end{cases}$$

is the absolute value of the largest possible payoff difference divided by $(1-e)$, there are $n = R$ retaliation stages, and

$$\begin{aligned} E &= [(1-e)-(1-e)^2]k + [(1-e)^2-(1-e)^3]p + \cdots \\ &\quad \cdots + [(1-e)^{R-1}-(1-e)^R]\big([1-\mathbf{1}_E(R)]p + \mathbf{1}_E(R)k\big) \\ &\approx e(\lfloor \tfrac{R}{2} \rfloor k + \lfloor \tfrac{R-1}{2} \rfloor p). \end{aligned}$$

We see that $\hat{B} = \frac{-1}{1-e}B_{P,r}|_{\pi_{D,r}=1}$ for odd numbers of retaliation stages and sensible values of the error rate e, as well as for a large range of parameter

[28] Obviously, both variants of the defecting type remain an ESS in the retaliation model for any strictly positive conformist bias. The latter is needed for the type in question to be stable against invasion by neutral drift by its also defecting counterpart.

values for an even number of retaliation stages including those we usually see in economic experiments.[29] We therefore focus our attention on this case. Note that the difference between equations (4) and (5) is merely in the largest possible payoff difference in the game, $-(1-e)\hat{B}$, the reciprocal of the payoff-normalising parameter β. Comparing equations (4) and (5), we directly see that for any number of additional retaliation stages n, $\hat{B}(n) > c + k = \hat{B}(0) \equiv \hat{B}^{base}$ will always hold, so that it will always require a lower conformist bias for accommodating punishers to be an ESS than for punishers in the baseline model. In other words, the introduction of retaliation opportunities may *enhance* the stabilisation of punishment. Furthermore, it is easy to show that $\hat{\alpha}_{P,a}(n+1) < \hat{\alpha}_{P,a}(n)$ always holds, such that a further increase in retaliation stages always leads to a higher likelihood of a cooperative outcome if punishers abstain from retaliation.

2. *retaliating punishers.* We already established that a player type setting out to invade a society consisting exclusively of punishers will not play a retaliating strategy with positive probability, given the debilitating effect on any individual within the majority will be very weak compared to the impact such behaviour will have on the invader's fitness. Since $B_{P,a} < B_{C,a}$ holds for any distribution of types $\boldsymbol{\pi}$, an optimally mixing invader will never include accommodating punishment in the support, either. From there, following a similar argument as in the baseline model we can directly derive the critical value of the conformist bias,

$$\hat{\alpha}_{P,r} = \begin{cases} \frac{ek+e[p+k]\Sigma_{r=1}^{R}(1-e)^r}{\hat{B}+ek+e[p+k]\Sigma_{r=1}^{R}(1-e)^r}, & c < (1-e)p \\ \\ \frac{c-(1-e)p+ek+e[p+k]\Sigma_{r=1}^{R}(1-e)^r}{\hat{B}+c-(1-e)p+ek+e[p+k]\Sigma_{r=1}^{R}(1-e)^r}, & \text{otherwise.} \end{cases} \tag{6}$$

Comparing equations (5) and (6) yields a confirmation of something we already know, namely that it is much easier to stabilise accommodating punishment than to stabilise the retaliating variant.[30] What will be the effect of introducing retaliation opportunities compared to the baseline scenario? We already know that the payoff-normalising parameter β decreases for $n > 0$, which we have seen to facilitate the stabilisation of punishment. Nevertheless, a comparison of (4) and (6) coupled with some straightforward algebra shows that, for any number of stages $n, n > 0$, we obtain $\hat{\alpha}_P^{base} < \hat{\alpha}_{P,r}(n)$. In other words, it needs a stronger conformist bias to ensure evolutionary stability of punishers in an environment where retaliation is possible if the punishers make use of that possibility. At the same time, $\hat{\alpha}_{P,r}(n)$ is not necessarily a monotonous function of n:

[29] Cf. footnote 27; even for $R = 7$ and with $k : p = 1 : 3$, we obtain a condition that is roughly equivalent to $c > (1/3)p + 2ep$, which is fulfilled for moderate error rates and a c-to-p ratio as in most experiments.

[30] To see that, note that the difference between $\hat{\alpha}_{P,a}$ and $\hat{\alpha}_{P,r}$ is simply that the same positive number is added to both the enumerator and the denominator of 5 to obtain 6.

while for $c < (1-e)p$, an even number of retaliation stages will always lead to a higher critical value of α than the next-lower odd number, the converse does not always hold and is highly dependent on the chosen parameters. For our parametrisation, we have $\hat{\alpha}_{P,r}(n+1) > \hat{\alpha}_{P,r}(n), \forall n < 4$, but the opposite holds for all n' such that $n' > 3$ and $\mathbf{1}_E(n') = 1$. For $c > (1-e)p$, no general statements can be made, as the result crucially depends on the c-to-p ratio. For the parameters used in this paper, the $\hat{\alpha}_{P,r}(n+1) > \hat{\alpha}_{P,r}(n), \forall n < 3$, and $\hat{\alpha}_{P,r}(n+1) < \hat{\alpha}_{P,r}(n)$, otherwise.

3. *accommodating pure cooperators.* In an environment without punishers, retaliators will always have the same payoff as non-retaliatiors. In the absence of any conformist bias, accommodating pure cooperators will be invaded by their retaliating counterpart by neutral drift. However, this will be prevented by any strictly positive conformist bias. Furthermore, it is easy to see that in this environment, an invading punisher of any variant will always have a lower expected payoff than a cooperator, such that, if accommodating pure cooperators are to be invaded, this will have to be done by any of the defector types. Analogously to the baseline model, we directly obtain

$$\hat{\alpha}_{C,a} = \frac{c}{\hat{B} + c}\,. \tag{7}$$

   Comparing equation (7) to equation (6), we see that, for sufficiently small values of the error frequency e, a reasonable number of retaliation stages R, and a wide range of parameter combinations (c, k, p) it will be easier to stabilise retaliating punishers than accommodating pure cooperators.[31] Nevertheless, the *a-priori* relationship between $\hat{\alpha}_{C,a}$ and $\hat{\alpha}_{P,r}$ is unclear. Having as little as five retaliation stages, an error rate of $e = 0.15$ and a c-to-p ratio of $2:3$ is enough to make the stabilisation of accommodating cooperators easier than that of retaliating punishers ($\hat{\alpha}_{C,a} = 0.167$ vs. $\hat{\alpha}_{P,r} = 0.170$).[32] For pure cooperators, we replicate a finding we already stated for the case of accommodating punishers: the higher the number of additional retaliation stages, the lower the requirement on the conformist bias for pure cooperation to be ES.

4. *retaliating pure cooperators.* For this group, the same holds as for accommodating pure cooperators, except for the fact that punishers invading in low numbers will obtain even lower payoffs than against the accommodating variant. Consequently, $\hat{\alpha}_{C,r} = \hat{\alpha}_{C,a}$ will hold, where $\hat{\alpha}_{C,a}$ is defined by equation (7).

Before we summarise the findings from our analysis and highlight some possible effects, let me shortly comment on the difference between the baseline and the retaliation models driving many of our above results. As we pointed out, the difference between equations (5) and (7), on the one hand, and (4), on the other, is merely in the largest possible payoff difference in the game, $(1-e)\hat{B}$, which is

[31] Cf. footnote 27.

[32] For comparison, under the same setting, $\hat{\alpha}_{P,a} = 0.015$.

the reciprocal of the payoff-normalising parameter β. This payoff difference may be interpreted as the reference intervall against which the types' actual payoff differences are evaluated. In a retaliative world, the largest potential payoff difference increases, as punishers may now face retaliative actions additionally to their own contribution and punishment expenses. Because of that, the types' normalised payoff differences become relatively smaller, and therefore, the payoff disadvantage faced by punishers to be outweighed by a conformist bias is relativised. In a sense, the finding that a cooperative outcome becomes more likely in a world where retaliation is an option but all players *abstain* from it bears a vague resemblance to the results from gift-exchange experiments that principals may do best *not* using their stick in a punishment world (cf. Fehr and Rockenbach, 2003).

Turning to the main results in terms of our analysis of evolutionary stability, we can state the following:

Result 2. As expected, retaliating punishers require a higher conformist bias to be ESS than their accommodating counterpart. For pure cooperators and defectors, the variant does not play a role in terms of the respective fixed point's evolutionary stability.

Result 3. The existence of retaliation opportunities makes the stabilisation of the accommodating variant of both pure cooperators and punishers easier compared to the model without such opportunities. Each additional stage lowers the critical value of the conformist bias for the respective type to be an ESS.

Result 4. Compared to punishment in the baseline game, for retaliating punishment to be ES a higher conformist bias is needed. The relationship between the critical value of α and the number of retaliation stages is non-monotonic and highly dependent on the c-to-p ratio.

Result 5. In many cases, the conformist bias needed to make retaliating punishment an ESS will be lower than that to stabilise pure cooperation. However, this need not always be the case.

Having talked about the evolutionary stability of the different player-types and having established the corresponding requirements on the conformist bias, we now want to illustrate a possible scenario that may come out of our dynamic system. For this illustrative purpose, we choose initial fractions of punishers that may seem excessively large when compared to common type distributions in the typical subject pool (Fischbacher et al., 2001, or Herrmann et al., 2008). Note however, that in public-goods experiments it is typically the *fraction of cooperative matching groups* determining the development of average contributions, rather than e.g. the contribution level of the least cooperative matching group. Hence, our focus on groups who *a priori* feature a greater disposition towards a cooperative outcome does not seem completely unwarranted *per se*. Whether the chosen type-distribution is excessively high even under these premises is a question we do not purport to answer empirically as our aim at this point merely is to illustrate possible *patterns*. Figure (3) depicts the type distributions we

obtain for a certain initial configuration under $\alpha = 0.1, e = 0.01$, and a) the baseline model, b) the retaliation model with $n = 1$ additional stage, and c) the retaliation model with $n = 4$.[33]

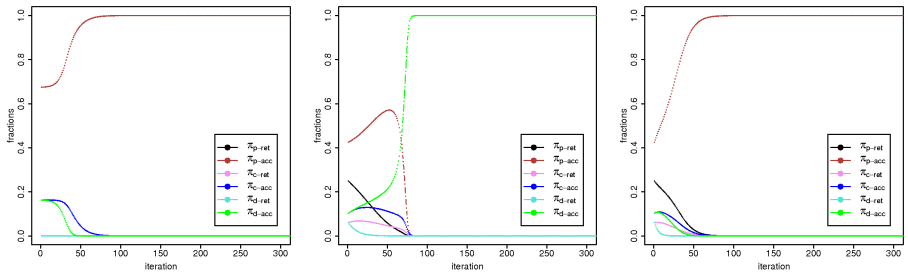

Figure 3: Evolution of types under a) the baseline model (left), b) the retaliation model with $n = 1$ (middle), and c) the retaliation model with $n = 4$ (right), for $\pi_P^0 = 0.675, \pi_C^0 = \pi_D^0 = 0.1625, \pi_{i,r}/\pi_i = 0.37, \forall i, \alpha = 0.1$ and $e = 0.01$.

What figure 3 provides is an example for an initial type distribution that would lead to the prediction of a cooperative outcome in both the baseline and the retaliation model with five retaliation stages, but to a non-cooperative-outcome prediction when there is only one such stage.

In figure 4 we illustrate the effect a small reduction in initial retaliativeness can have. Reducing retaliativeness in the model with $n = 1$ by one percentage point, we cross the border of basins of attraction and end up with a cooperative outcome instead of the defecting outcome depicted in figure 3. Both figures cannot but give examples of potential patterns that may emerge, and yet they show that the pattern of dynamics observed in the experimental studies are not inherently inconsistent with what an evolutionary model would predict. In our next section, we elaborate on this issue in a little more detail, discussing the theoretical results obtained against the backdrop of the experimental findings of punishment studies like Fehr and Gächter (2000, 2002) and of counterpunishment studies like Nikiforakis (2008). We further put our results into the context of earlier evolutionary models of cooperation based on punishment before concluding in section 5. In that section, we shortly recapitulate our results and address the two main questions of this paper: (i) can we still explain cooperation by refering to punishment mechanisms if we accept that counterpunishment cannot be excluded, and (ii) if we can, will retaliation have to play a role in future models of cooperation?

[33] The numbers of stages were chosen as to reflect the experimental setups of Nikiforakis (2008) and Denant-Boemont et al. (2007). Similar type-evolution patterns can be found for other values of n, α, and e.

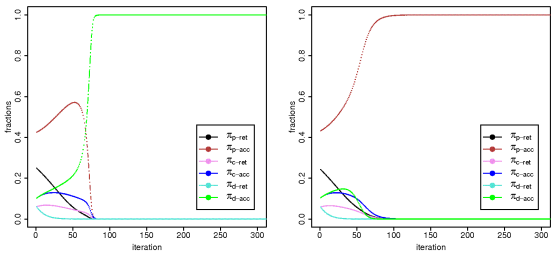

Figure 4: Evolution of types under the retaliation model with $n = 1$ for an initial retaliativeness level $\pi_{i,r}/\pi_i = 0.37, \forall i$ (left) and b) for $\pi_{i,r}/\pi_i = 0.36, \forall i$ (right), for $\pi_P^0 = 0.675, \pi_C^0 = \pi_D^0 = 0.1625, \alpha = 0.1$ and $e = 0.01$.

4. Discussion

How do our results relate to the findings of experimental studies? Studies like Fehr and Gächter (2000, 2002) have shown that single-directional punishment is able to maintain high contribution levels and even increase them. This closely corresponds to our result 1: punishment can stabilise cooperation in a model without counterpunishment even for low values of a conformist bias. In experimental studies allowing for retaliation, results depend on the experimental setup. While in Denant-Boemont et al. (2007, DNM) and Nikiforakis (2008), the introduction of a single additional punishment stage with a restriction to retaliative punishment leads to an elimination of any contribution-enhancing effect of punishment, the former show that this breakdown of cooperation is at least in part due to the restrictive assumption of a *single*, *retaliative* stage. In figure 3, we illustrated how these results may come about: under a range of initial type distributions, a cooperative outcome is observed for a single punishment stage as well as multiple retaliative stages, but not for a single counterpunishment stage.

Still, experimental results exhibit a large variance: in an unreported experiment, Kube et al. (2009) employ the same design as Nikiforakis (2008) but do not find a decline in average contributions; Nikiforakis and Engelmann (2008, NE) introduce an endogenous number of punishment stages and observe a cooperation level that is as high as in a treatment without opportunities for retaliation, which contrasts with the earlier findings of DNM.[34] In our view, the most likely explanation for these discrepancies is a subject-pool difference. As has been illustrated in figure 4, differences in players' retaliativeness may account for a shift in the expected outcome from cooperation to defection and *vice versa*. A possible reading of our results could see the elimination of the positive effects of punishment observed in the earlier counterpunishment studies stem from so-

[34] The reported findings by DNM and NE were obtained using a partner-matching with changing IDs; Nicklisch and Wolff (2009) find cooperation-levels similar to those of NE for a stranger-matching.

ciety consisting of sufficiently retaliative individuals, while the subjects used by the latter studies may have been more tolerant ('accommodating'). This difference in the subject pools' average responsiveness to punishment need not be large if the initial type distribution is close enough to the corresponding border of attraction.

With respect to the evolutionary literature, our result 1 discussed above confirms the results of earlier models without counterpunishment opportunities, such as Henrich and Boyd (2001) or Boyd et al. (2003). Incorporating retaliation, we find that the level of conformist learning required to make punishment an evolutionary stable strategy is higher compared to the model without counterpunishment if players' retaliativeness is high (result 4), but lower if retaliators are rare (result 3). Under the latter scenario, a purely cooperative strategy is also more likely to be stabilised than in the baseline model (results 2 and 3). A higher number of retaliation stages plays a facilitating role for the stability of purely cooperative strategies as well as accommodating punishment, due to the worsened threat by the consequences of a battle of punishment and counterpunishment. Finally, we have seen that for some parameters it may be harder to stabilise retaliative punishment than pure cooperation (result 5).

5. Conclusion

One of the explanations considered most often for the puzzle of the high degree of human cooperation has been the existence of punishment opportunities in conjunction with the presence of strong reciprocators. In fact, this seems to be the focal explanation for cooperation in experimental work on social-dilemmas with no repeated interaction between the same subjects under an anonymous setting. However, recent laboratory studies like Denant-Boemont et al. (2007) and Nikiforakis (2008) allowed for retaliative punishment, casting serious doubt on the hypothesis' explanatory power. While the cooperation-adverse results of Nikiforakis (2008) seem to hinge on the restriction to a single, retaliation-only stage, it remains unclear whether in a multiple-punishment-stage setting cooperation can be restored to its full degree. Experimental studies have led to contradictory findings on the question of whether the high degree of cooperation among humans can still be explained by the possibility to punish defecting society members (e.g., Nikiforakis and Engelmann, 2008), or whether this is a rather unlikely explanation (Denant-Boemont et al., 2007). To our knowledge, the present study is the first to address this question theoretically, using an evolutionary model of cooperation with multiple punishment and counterpunishment stages. We find that our model can account for the breakdown of cooperation under a single retaliation stage as well as for its 'restoration' under multiple retaliation stages. Furthermore, our results suggest that the degree of 'restoration' crucially depends on the population's retaliativeness level, suggesting that the contradictory laboratory findings could easily be explained by hardly noticeable subject-pool differences.

From a theoretic perspective, our findings suggest that not to account for retaliation may not be an overly restrictive assumption when explaining cooper-

ation *as a phenomenon*: counterpunishment may change a society's probability of attaining a cooperative outcome, it may do so for both better and worse; at the same time, it generally does not render the punishing fixed point unattainable, unless the level of conformist learning is very low. Therefore, if a conformist bias is a sensible descriptor of a part of social learning processes, the introduction of counterpunishment opportunities does not change the qualitative results of a punishment model of cooperation. In other words, if conformist learning is a non-negligible part of reality, abstracting from retaliation in models of cooperation induced by punishment may seem a sensible assumption to make – as long as we do not want to make quantitative predictions about the probability of a cooperative outcome.

References

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

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

Andrés Guzmán, Ricardo, Carlos Rodriguez-Sickert, and Robert Rowthorn (2007): When in Rome, do as the Romans do: the coevolution of altruistic punishment, conformist learning, and cooperation. *Evolution and Human Behavior* 28, 112-117.

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

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

Boyd, Robert, Herbert Gintis, Samuel Bowles, and Peter J. Richerson (2003): The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences of the United States of America* 100, 3531-3535.

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

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.

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.

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

Decker, Torsten, Andreas Stiehler, and Martin Strobel (2003): A Comparison of Punishment Rules in Repeated Public Good Games. *Journal of Conflict Resolution* 47(6), 751-772.

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

Dreber, Anna, David G. Rand, Drew Fudenberg, and Martin A. Nowak (2008): Winners don't punish. *Nature* 452, 348-351.

REFERENCES

Egas, Martijn, and Arno Riedl (2008): The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B: Biological Sciences* 275, 871-878.

Fehr, Ernst, and Simon Gächter (2002): Altruistic punishment in humans. *Nature* 415, 137-150.

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

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.

Fehr, Ernst, and Bettina Rockenbach (2003): Detrimental Effects of Sanctions on Human Altruism. *Nature* 422, 137-140.

Fischbacher, Urs, Simon Gächter, and Ernst Fehr (2001): Are People Conditionally Cooperative? Evidence from a Public Goods Experiment. *Economic Letters* 71(3), 397-404.

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.

Gächter, Simon, Elke Renner, and Martin Sefton (2008): The Long-Run Benefits of Punishment. *Science* 322, 1510.

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

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

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

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

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

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.

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.

REFERENCES

Henrich, Joseph, and Natalie Henrich (2006): Culture, evolution and the puzzle of human cooperation. *Cognitive Systems Research, Cognition, Joint Action and Collective Intentionality* 7, 220-245.

Henrich, Joseph, and Richard McElreath (2003): The evolution of cultural evolution. *Evolutionary Anthropology: Issues, News, and Reviews* 12, 123-135.

Henrich, Joseph, Richard McElreath, Abigail Barr, Jean Ensminger, Clark Barrett, Alexander Bolyanatz, Juan C. Cardenas, Michael Gurven, Edwins Gwako, Natalie Henrich, Carolyn Lesorogol, Frank Marlowe, David Tracer, and John Ziker (2006): Costly Punishment Across Human Societies. *Science* 312, 1767-1770.

Herrmann, Benedikt, Christian Thöni, and Simon Gächter (2008): Antisocial Punishment Across Societies. *Science* 319, 1362-1367.

Hruschka, Daniel J., and Joseph Henrich (2006): Friendship, cliquishness, and the emergence of cooperation. *Journal of Theoretical Biology* 239, 1-15.

Ledyard, John O. (1995): Public Goods: A Survey of Experimental Research. In: John Kagel and Alvin Roth (eds.): *Handbook of Experimental Economics.* Princeton, NJ: Princeton University Press.

McKenzie Alexander, Jason (2005): The Evolutionary Foundations of Strong Reciprocity. *Analyse & Kritik* 27, 106-112.

Kube, Sebastian, Andreas Nicklisch, and Christoph Engel (2009): personal communication.

Nicklisch, Andreas, and Irenaeus Wolff (2009): Governing the Sword: the Effects of a Basic Institution in a Social Dilemma. *Working Paper.*

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

Nikiforakis, Nikos, and Dirk Engelmann (2008): Feuds in the Laboratory? A Social Dilemma Experiment. *Working Paper.*

Nikiforakis, Nikos, and Hans-Theo Normann (2008): A Comparative Statics Analysis of Punishment in Public-good Experiments. *Experimental Economics* 11, 358-369.

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

Nowak, Martin A., and Karl Sigmund (2005): Evolution of indirect reciprocity. *Nature* 437, 1291-1298.

Ostrom, Elinor (2000): Collective Action and the Evolution of Social Norms. *Journal of Economic Perspectives* 14(3), 137-158.

REFERENCES

Ostrom, Elinor, James M. Walker, and Roy Gardner (1992): Covenants With and Without a Sword: Self-Governance is Possible. *The American Political Science Review* 86, 404-417.

de Quervain, Dominique J. F., Urs Fischbacher, Valerie Treyer, Melanie Schellhammer, Ulrich Schnyder, Alfred Buck, and Ernst Fehr (2004): The Neural Basis of Altruistic Punishment. *Science* 305, 1254-1258.

Richerson, Peter J., and Robert Boyd (2005): *Not by Genes Alone* Chicago: The University of Chicago Press.

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

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

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

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

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

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

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

Zafar, Basit (2009): An Experimental Investigation of Why Individuals Conform. *Federal Reserve Bank of New York Staff Reports* 365.

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