
# NOTA DI LAVORO

64.2011

## Punish and Perish?

By **Angelo Antoci**, DEIR, University of Sassari, Italy

**Luca Zarri**, Economics Department, University of Verona, Italy

Economy and Society
Editor: Giuseppe Sammarco

# Punish and Perish?

By Angelo Antoci, DEIR, University of Sassari, Italy
Luca Zarri, Economics Department, University of Verona, Italy

## Summary

The evolution of large-scale cooperation among genetic strangers is a fundamental unanswered question in the social sciences. Behavioral economics has persuasively shown that so called 'strong reciprocity' plays a key role in accounting for the endogenous enforcement of cooperation. Insofar as strongly reciprocal players are willing to costly sanction defectors, cooperation flourishes. However, experimental evidence unambiguously indicates that not only defection and strong reciprocity, but also unconditional cooperation is a quantitatively important behavioral attitude. By referring to a prisoner's dilemma framework where punishment ('stick') and rewarding ('carrot') options are available, here we show analytically that the presence of *cooperators who don't punish* in the population makes altruistic punishment evolutionarily weak. We show that cooperation breaks down and strong reciprocity is *maladaptive* if costly punishment means 'punishing defectors' and, even more so, if it is coupled with costly rewarding of cooperators. In contrast, punishers don't perish if cooperators, far from being rewarded, are *sanctioned*. These results, based on an extended notion of strong reciprocity, challenge evolutionary explanations of cooperation that overlook the 'dark side' of altruistic behavior.

**Keywords:** Cooperation, Strong Reciprocity, Altruistic Punishment, Altruistic Rewarding, Heterogeneous Types

**JEL Classification:** C7, D7, Z1

*Address for correspondence:*

Angelo Antoci
University of Sassari
via Muroni 25
07100 Sassari
Italy
E-mail: angelo.antoci@virgilio.it

# Punish and Perish?

Angelo Antoci, DEIR, University of Sassari, via Muroni 25, 07100 Sassari (Italy).

E-mail: angelo.antoci@virgilio.it

Luca Zarri, Economics Department, University of Verona, Viale Università 3, 37129 Verona (Italy).
E-mail: luca.zarri@univr.it

**Abstract**

The evolution of large-scale cooperation among genetic strangers is a fundamental unanswered question in the social sciences. Behavioral economics has persuasively shown that so called 'strong reciprocity' plays a key role in accounting for the endogenous enforcement of cooperation. Insofar as strongly reciprocal players are willing to costly sanction defectors, cooperation flourishes. However, experimental evidence unambiguously indicates that not only defection and strong reciprocity, but also unconditional cooperation is a quantitatively important behavioral attitude. By referring to a prisoner's dilemma framework where punishment ('stick') and rewarding ('carrot') options are available, here we show analytically that that the presence of *cooperators who don't punish* in the population makes altruistic punishment evolutionarily weak. We show that cooperation breaks down and strong reciprocity is *maladaptive* if costly punishment means 'punishing defectors' and, even more so, if it is coupled with costly rewarding of cooperators. In contrast, punishers don't perish if cooperators, far from being rewarded, are *sanctioned*. These results, based on an extended notion of strong reciprocity, challenge evolutionary explanations of cooperation that overlook the 'dark side' of altruistic behavior.

**Key words**: Cooperation; Strong Reciprocity; Altruistic Punishment; Altruistic Rewarding; Heterogeneous Types.

**JEL Classification:** C7; D7; Z1.

## 1. Introduction

Even though the evolution of cooperation among humans has been extensively studied in the last decades (see e.g. the famous work by Axelrod, 1984), a crucial question remains largely open: how can large-scale cooperation endogenously emerge and be successfully sustained over time? Recent research has persuasively argued that invoking explanations based on more or less sophisticated forms of 'enlightened self-interest' alone, such as kin selection (Hamilton, 1964), repeated encounters (Fudenberg and Maskin, 1986), and reputation formation, is not sufficient for accounting for the evidence available about the relevance of cooperation within several significant human contexts where collective action problems naturally arise but interactions involve genetically unrelated individuals. As both theoretical contributions and empirical evidence confirm, insofar as altruists are grouped together and mainly interact among themselves, within a neighborhood structure (Eshel et al., 1998), exploitation on the part of free riders can be prevented by restricting access to the gains from cooperation. Unlike these studies, we depart from close-knit parochial communities and test the survival potential of pro-social behavior within a more 'hostile' environment where neither group selection nor assortative interactions are allowed, and develop an evolutionary game-theoretical analysis aimed at investigating the diffusion of cooperation when exogenous enforcement devices are not available.

In recent years, a growing body of experimental evidence has convincingly shown that so called 'strong reciprocity' is a powerful device for the enforcement of cooperation, despite the presence of large proportions of selfish subjects (Fehr and Gächter, 2000; Gintis et al., 2005; Gächter and Herrmann, 2010). The key feature of strong reciprocators is their willingness to incur costs in order to conditionally cooperate and punish non-cooperators. However, in a lively interdisciplinary debate currently involving economists, biologists and social psychologists (see on this Fehr and Henrich, 2003), critics argue that strong reciprocity is *maladaptive*, in the sense that it is evolutionarily weak and has no adaptive power (Dreber et al., 2008).

Hence, the following question naturally arises: can strong reciprocity survive and favor the enforcement of cooperation, within a behaviorally heterogeneous population in which *also non-reciprocating* players are involved? The existence of heterogeneous types is being increasingly confirmed by experimental research[1]. In particular, available lab evidence indicates that (a) a significant proportion are unconditionally cooperative (e.g. systematically cooperate in the prisoner's dilemma or make positive contributions in public goods or dictator games) and (b) a significant proportion of subjects are self-interested and tend to free ride on others'

[1] Fischbacher and Gächter (2010), by means of a new methodological strategy, both account for the existence of types in the lab and, through a direct test of the role of social preferences in voluntary cooperation, show that a large part of the dynamics of free riding is explained by the interaction of heterogeneous types.

generosity (by defecting from the outset). Further, (c) most subjects who act neither purely selfishly nor simply altruistically seem to be strong reciprocators (Carpenter et al., 2009).

However, it is worth pointing out that strong reciprocity is often viewed as something more than costly punishment of non-cooperators: in many existing works on the theme, a strongly reciprocal player is generically defined as a person who is willing to bear costs to be kind to those who are being kind (by cooperating and rewarding them; strong positive reciprocity) *and* to be mean to those who are being mean (by defecting and punishing them; strong negative reciprocity; Fehr et al., 2002). A relevant problem with this definition is that it takes for granted that if a person is willing to be kind to those who are being kind, she will also be mean to those who are being mean (or viceversa). By contrast, several experimental papers (see e.g. Abbink et al., 2000; Offerman, 2002; Reuben and van Winden, 2010) show that strong positive reciprocity need not be the flip-side of strong negative reciprocity. Moreover, a further extension of the notion of strong reciprocity is in order as some new studies interestingly reveal that punishers target their sanctions not only to defectors but also, to a significant extent, to other cooperators (Herrmann et al., 2008; Goette et al., 2010; Abbink et al., 2010; Gächter and Herrmann, 2010). This suggests that, on both conceptual and empirical grounds, strong reciprocity has a plural nature. In the light of this, in this paper we decompose such behavioral attitude by introducing a taxonomy of strongly reciprocal players. Next, we comparatively lay out the evolutionary foundations of the varieties of strong reciprocity we identify within a behaviorally heterogeneous social environment where also unconditional defectors and unconditional cooperators are initially present. This will allow us to explore – to our knowledge for the first time – the different medium-long run implications which can be drawn, for society at large, depending on the variety of strong reciprocity one refers to.

The structure of the remainder of this paper is as follows. Section 2 illustrates the main features of the analytical model. Section 3 analyzes social dynamics for each of the variants of strong reciprocity we investigate, and contains our basic mathematical results. Section 4 discusses the main findings and concludes.

## 2. The Prisoner's Dilemma with Carrot and Stick

We consider a large-scale population of individuals enjoying the benefits of a given collective (i.e. non-rival and non-excludable) good. In this society, three player types are initially present: *Unconditional Defectors* (*UD*s, hereafter 'defectors' only, for simplicity), *Unconditional Cooperators* (*UC*s, hereafter 'cooperators') and *Strong Reciprocators* (*SR*s). The existence and quantitative relevance of these behavioral types is being increasingly confirmed in laboratory environments, within the framework of prisoner's dilemma and public

good game experiments (Fehr and Fischbacher, 2005; Fischbacher and Gächter, 2010; Ones and Putterman, 2007; Camera and Casari, 2009; Carpenter et al., 2009). As we anticipated above, we take the inherently plural nature of strong reciprocity into account and model it by introducing a taxonomy of *SR* types. Infinitely many random encounters occur between two individuals at a time and, whenever two players meet, their behavior affects each other's enjoyment of the collective good. Besides, type recognition holds: players are supposed to be able to identify their co-player's type in each pairwise matching[2]. This feature of the model is in common with models of good standing (Panchanathan and Boyd, 2004) as well as previous evolutionary work on altruistic punishment (Fowler, 2005). In each matching, we assume that the material consequences for the players are captured by a two-stage Prisoner's Dilemma with Carrot and Stick (*PDCS*) game. In the first stage (the *Cooperation stage*), the material consequences depend on players' choices between 'Cooperate' (*C*) and 'Defect' (*D*) only. Hence, each matching between two individuals will produce one of the following four outcomes: (*D*, *D*), (*C*, *C*), (*C*, *D*), (*D*, *C*). We suppose that, from the viewpoint of the individual player, the material consequences of these four possible outcomes have the structure of the prisoner's dilemma, with $\gamma > \alpha > \beta > \delta$ (see the left side of Table 1). Mutual cooperation Pareto-dominates mutual defection and free riding, by exploiting the non-excludability of the good to be provided and the opponent's cooperation, is the most individually rewarding outcome, in material terms. Players behave according to their type. Hence, while *UCs* play *C* and *UDs* play *D* in each matching (regardless of the opponent's type), *SRs* play *C* when matched with another *SR* or with a *UC* player, and play *D* when matched with a *UD* player (see the right side of Table 1).

| | *C* | *D* |
|---|---|---|
| *C* | $\alpha, \alpha$ | $\delta, \gamma$ |
| *D* | $\gamma, \delta$ | $\beta, \beta$ |

*PD game*

| | *UC* | *SR* | *UD* |
|---|---|---|---|
| *UC* | $\alpha, \alpha$ | $\alpha, \alpha$ | $\delta, \gamma$ |
| *SR* | $\alpha, \alpha$ | $\alpha, \alpha$ | $\beta, \beta$ |
| *UD* | $\gamma, \delta$ | $\beta, \beta$ | $\beta, \beta$ |

*Players' material payoffs*

**Table 1:** *PD game and material payoffs (Stage 1)*

In the second stage (the *Punishment/Reward stage*), players have to choose among Punish, *P* ('stick'), Reward, *R* ('carrot') and Neither, *N* (that is, abstaining from both punishment and rewarding). Each *SR* chooses *N* if matched with a player of the same type. Further, we suppose that while both *UCs* and *UDs*

[2] Though throughout the paper we retain this information assumption, it is worth pointing out that most of our results can be obtained also if we suppose that players can observe their opponent's type only ex post, that is after playing the first stage of the material game.

systematically abstain from punishing and/or rewarding others, *SR* types are classified according to their choices (*P*, *R* or *N*) when matched with *UCs* and *UDs* (see the left side of Table 2). In particular, we separately consider five types of players who act identically in the first stage – that is, they play *C* (resp., *D*) when matched with either a *UC* or another *SR* (resp., a *UD)* – but differ as to their strategic choice in the second stage. In particular, as the left side of Table 2 shows, we specifically focus on (1) strong negative reciprocators (*SNR*s), who only punish defectors; (2) strong positive reciprocators (*SPR*s), who only reward cooperators; (3) symmetric strong reciprocators (*SSR*s), who both punish defectors and reward cooperators; (4) punishers of non-punishing cooperators (*PNP*s), who only punish cooperators and, finally, (5) hyper-strong negative reciprocators (*HSNR*s), who punish both cooperators and defectors[3]. The matrix on the right of Table 2 provides us with the material payoffs at stage 2, where $\varepsilon$ = *cost of being punished*, $\lambda$ = *cost of punishing*, $\pi$ = *cost of rewarding*, $\eta$ = *benefit from being rewarded* and $\lambda, \varepsilon, \pi, \eta$ are strictly positive parameters. This means that, as it is often true both in naturally occurring environments (Sethi and Somanathan, 1996) as well as in laboratory experiments (Gächter and Herrmann, 2010), punishing (resp., rewarding) is a costly activity for the punisher (resp., rewarder).

| | *UC* | *UD* |
|---|---|---|
| *SNR* | *N* | *P* |
| *SPR* | *R* | *N* |
| *SSR* | *R* | *P* |
| *PNP* | *P* | *N* |
| *HSNR* | *P* | *P* |

*SRs' classification*

| | *UC* | *UD* |
|---|---|---|
| *SNR* | 0,0 | $-\lambda, -\varepsilon$ |
| *SPR* | $-\pi, \eta$ | 0,0 |
| *SSR* | $-\pi, \eta$ | $-\lambda, -\varepsilon$ |
| *PNP* | $-\lambda, -\varepsilon$ | 0,0 |
| *HSNR* | $-\lambda, -\varepsilon$ | $-\lambda, -\varepsilon$ |

*SRs' material payoffs*

**Table 2:** *SRs' classification and material payoffs (Stage 2)*

We claim that the two-stage structure of the *PDCS* allows us to go beyond a further limitation which characterizes existing studies on strong reciprocity, that is their inability to sharply distinguish between *implicit* and *explicit* forms of rewarding and punishment. With regard to rewarding, one may argue that, for example in a prisoner's dilemma game, deciding to cooperate with a cooperator entails in itself sacrificing resources to be kind towards (i.e. to reward) a person being kind (strong positive reciprocity), since the same

[3] Though our analysis includes forms of strong reciprocity leading to punishment of cooperators, it is worth pointing out that in this paper we leave aside the interesting phenomenon of defectors punishing cooperators (the so called 'antisocial punishment'). For a recent theoretical work on the impact of anti-social punishment on the evolution of cooperation, see Rand et al. (2010).

person would have obtained a larger material benefit by defecting (rather than by cooperating) with a cooperating player. Analogously, defection can be seen as an implicit means of punishing defectors. The Folk Theorem literature provides us with two famous examples of implicit punishment via defection such as Tit-for-Tat and the Grim Trigger strategy. By contrast, the structure of the *PDCS* allows us to incorporate two levels of punishment and rewarding into the analysis, so that strong reciprocity turns out to be a behavioral attitude characterized by *both* conditional niceness (i.e. willingness to cooperate with cooperators and to defect with defectors) *and* costly acts of punishment and/or rewarding[4].

## 3. The evolutionary game-theoretical model: do punishers perish?

As we made clear in the previous section, in our evolutionary game-theoretical model player types prescribe the behavioral patterns which, via pairwise matchings, determine specific material consequences. In turn, such material consequences drive social evolution, in the sense that the types which turn out to be more rewarding – in *material* terms – are imitated and, by replicating faster, manage to spread over at the expense of less rewarding ones. Time is continuous and the population is modelled as a continuum of players. As far as pairwise matchings are concerned, the material game that individuals play is the previously described two-stage *PDCS* game. We represent the state of the population of individuals by the vector $x = (x_1, x_2, x_3) \in R^3$, where $x_1$, $x_2$ and $x_3$ indicate the shares of individuals of the types *UC*, *SR* and *UD*, respectively. Thus $x_i \geq 0$, for all *i*, and $\sum_i x_i = 1$; so *x* belongs to the 2-dimensional simplex S (see Figure 1). Let us indicate by *A* the payoff matrix of the *PD*CS game associated with the material payoffs related to both stage 1 and stage 2 (see Tables 1 and 2), whose entries $a_{ij}$ depend on the specific *SR* type considered and represent the row player's payoffs corresponding to each pairwise interaction, in a *UC*-*SR*-*UD* population:

$$A = \begin{array}{c|ccc} & UC & SR & UD \\ UC & a_{11} & a_{12} & a_{13} \\ SR & a_{21} & a_{22} & a_{23} \\ UD & a_{31} & a_{32} & a_{33} \end{array} \qquad (1)$$

[4] The distinction between implicit and explicit punishment and rewarding is in line with experimental evidence, indicating that subjects often behave differently according to whether they are provided or not with explicit opportunities to directly target their sanctions and/or rewards towards other players (Fehr and Fischbacher, 2003; 2005).

Given the pairwise random matching structure of the game, the (expected) material payoffs for *UC*s, *SR*s and *UD*s are, respectively:

$$\begin{aligned}\Pi_{UC} &= a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ \Pi_{SR} &= a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \\ \Pi_{UD} &= a_{31}x_1 + a_{32}x_2 + a_{33}x_3\end{aligned}$$

Following Taylor and Jonker (1978), we assume that the growth rates $\dot{x}_i / x_i = (dx_i / dt) / x_i$ of the shares are given by the well-known replicator equations (see also Weibull, 1995):

$$\begin{aligned}\dot{x}_1 &= x_1\left(\Pi_{UC} - \overline{\Pi}\right) \\ \dot{x}_2 &= x_2\left(\Pi_{SR} - \overline{\Pi}\right) \\ \dot{x}_3 &= x_3\left(\Pi_{UD} - \overline{\Pi}\right)\end{aligned} \qquad (2)$$

where:

$$\overline{\Pi} = x_1\Pi_{UC} + x_2\Pi_{SR} + x_3\Pi_{UD}$$

represents the population-wide average payoff.

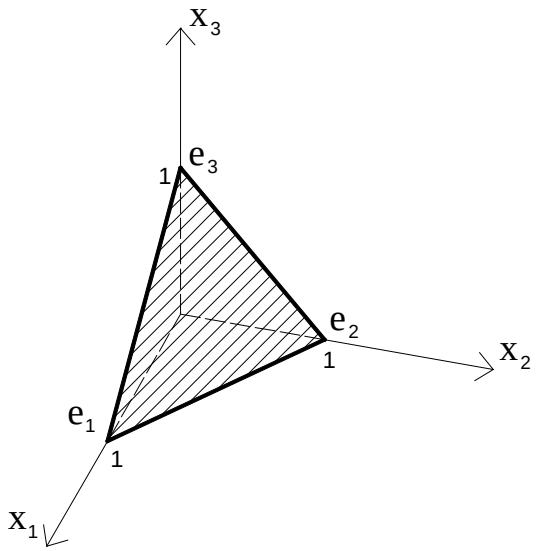



**Figure 1:** *The* 2*-dimensional simplex S*

Our formalization allows us to directly draw implications about the social dynamics taking place within large-scale three-type populations in which cooperators and defectors initially coexist with strong negative reciprocators (*SNR*s), strong positive reciprocators (*SPR*s), symmetric strong reciprocators (*SSR*s), punishers of non-punishing cooperators (*PNP*s) and hyper-strong negative reciprocators (*HSNR*s), respectively.

The dynamic system (2) is analyzed in the Mathematical Appendix by using the classification due to Bomze (1983) for replicator equations. In the following subsections we illustrate the basic features of dynamics generated by (2) by separately focusing on each of the five varieties of strong reciprocity under exam.

In Figures 2-6, attractive stationary states are indicated by full dots, repulsive ones by open dots and saddle points by drawing their stable and unstable branches. The vertices $e_1 = (1,0,0)$, $e_2 = (0,1,0)$ and $e_3 = (0,0,1)$ of the simplex S (see Figure 1), which represent respectively the states where only types *UC*, *SR* and *UD* are present in the population, are indicated by *UC*, *SR* and *UD*. These states are always stationary states under replicator dynamics.

***Altruistic Punishers***

Figure 2 illustrates the dynamics emerging when *SR*s display a willingness to costly punish defectors only (*SNR*), consistently with many laboratory studies (see Gintis et al., 2005), where a sizeable proportion of *SNR*s is identified. In such a context, the payoff matrix A becomes:

$$A = \begin{array}{c|ccc} & UC & SNR & UD \\ UC & \alpha & \alpha & \delta \\ SNR & \alpha & \alpha & \beta - \lambda \\ UD & \gamma & \beta - \varepsilon & \beta \end{array}$$

Observe that a *UC*-*UD*-*SNR* population may end up either in a ‘bad’ stationary state (the vertex *UD*), where cooperators and strong reciprocators perish and all players are defectors, or in a ‘good’ stationary state belonging to the edge joining the vertices *UC* and *SNR* (every point of such an edge is a stationary state) where defectors perish, with positive proportions of cooperators and strong reciprocators. However, the latter evolutionary outcome is fragile and the maintenance of cooperation may be jeopardized: if the share of *SNR*s falls below a certain threshold in the polymorphic stationary states of the edge *UC*-*SNR*, such polymorphic configurations can be invaded by defectors. This result is in line with past evolutionary work (Sethi and Somanathan, 1996) and experimental evidence (Carpenter et al., 2004) revealing that when ‘sufficiently

many' punishers are initially present, free riders are likely to be matched with agents reducing their payoffs, so that the former will be driven out of the population. At that point, since there will be no selection pressure against punishing players, the population shares stabilize. In such a case, a polymorphism with a positive proportion of two pro-social behavioral types (cooperators and (a high enough number of) punishers) and universal cooperation prevails. In our analysis, we also find that, other things being equal, as defectors' costs of being punished increase, the basin of attraction of the vertex *UD* becomes smaller (see the Mathematical Appendix). This can be seen as an evolutionary confirmation of what Sethi and Somanathan (1996) refer to as the centrepiece of economic reasoning, that is "the tendency of human behaviour to adjust in response to persistent differential in material incentives".

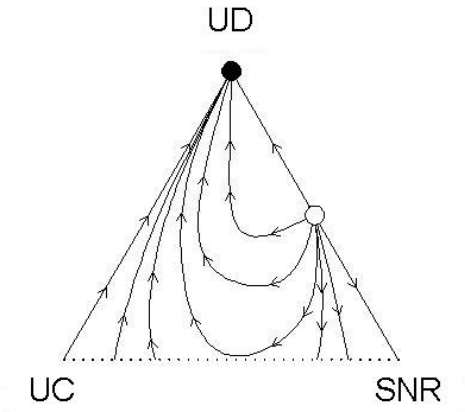



**Figure 2:** *Social dynamics in a population of Cooperators, Defectors and Altruistic Punishers*

***Altruistic Rewarders***

This case represents a scenario where cooperators and defectors coexist with players driven by *SPR*, that is conditional cooperators who are willing to incur costs to reward cooperators (altruistic rewarding), but abstain from punishing defectors, unlike agents driven by *SNR*. In such case, the payoff matrix A is given by:

$$A = \begin{array}{c|ccc} & UC & SPR & UD \\ UC & \alpha & \alpha+\eta & \delta \\ SPR & \alpha-\pi & \alpha & \beta \\ UD & \gamma & \beta & \beta \end{array}$$

In their public goods experiment on endogenous institutional choice (carrot vs. stick), Sutter et al. (2010) find that subjects typically vote for the reward option. In this case, our analytical model shows that the three types

coexist in positive, permanently fluctuating proportions (Figure 3). Such a dynamics qualitatively resembles one of the findings obtained in the well-known evolutionary paper on indirect reciprocity by Nowak and Sigmund (1998), where it is shown that long-term simulations that incorporate mutations usually do not converge to a simple equilibrium distribution of strategies, but display endless cycles, with defectors, discriminators and cooperators. This is the only coexistence outcome we obtain (though we do not get an attractive stationary state with coexistence), with reference to both behavioral types (as selfish and non-selfish players coexist) and behavioral outcomes (as we observe both cooperation and defection, within the overall population). By contrast, all the other four varieties of strong reciprocity we investigate lead to the survival of either selfish or non-selfish (i.e. cooperators and/or strong reciprocators) players only, which either universal defection or universal cooperation is associated with.

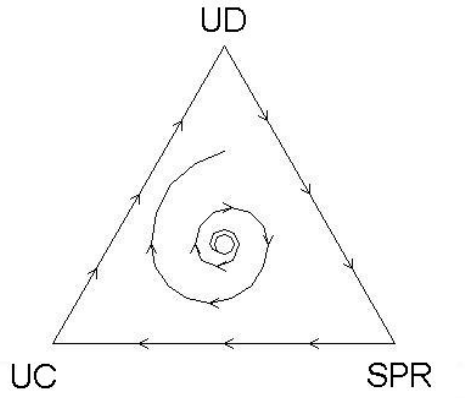



**Figure 3:** *Social dynamics in a population of Cooperators, Defectors and Altruistic Rewarders*

***Players driven by Symmetric Strong Reciprocity***

Let us now consider the dynamics associated with the case in which cooperators and defectors initially coexist with conditional cooperators displaying *SSR*, that is the combination of altruistic punishment (punishment of defectors) and altruistic rewarding (rewarding of cooperators; Fehr et al., 2002). In this case, the payoff matrix A is:

$$A = \begin{array}{c|ccc} & UC & SSR & UD \\ UC & \alpha & \alpha+\eta & \delta \\ SSR & \alpha-\pi & \alpha & \beta-\lambda \\ UD & \gamma & \beta-\varepsilon & \beta \end{array}$$

Here, we find that the stationary state *UD*, where all players are defectors, is a global attractor in the interior of the simplex S (Figure 4). The strong result we obtain is that now complete free riding prevails regardless of the proportion of non-selfish players (*SR*s and *UC*s) initially present in the population. The intuition, in a nutshell, is that altruistic rewarding 'crowds-out' altruistic punishment. What happens in this case resembles the well-known dynamics characterizing a classic *prey-predator* model, but within a cultural evolution framework in which different cultural orientations compete with one another and evolution is driven by material payoffs. Within the behaviorally heterogeneous framework under study, *SSR* is maladaptive due to the key negative role played by the group of cooperators, as such players, by so doing, make themselves vulnerable and exploitable on the part of *UD*s, so favoring their evolutionary success. As we have seen by analyzing *SNR*, such an unpleasant social outcome can be prevented – provided that 'sufficiently many' punishers are initially present –, as *SR*s in that case abstain from rewarding cooperators. By contrast, with *SSR* universal defection prevails regardless of the initial share of *SR*s in the population. Since selection favors second-order free riders, strong reciprocity declines and eventually first-order free riders take over. This is a crucial point which, though speculatively made (Panchanathan and Boyd, 2004) or investigated by means of exploratory simulations (Fehr and Fischbacher, 2003), had not received specific attention so far at the analytical level.

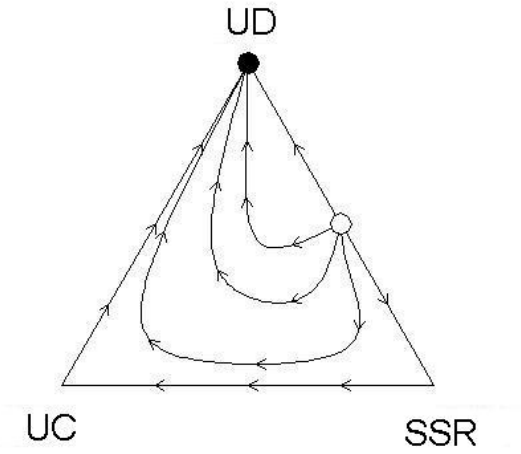


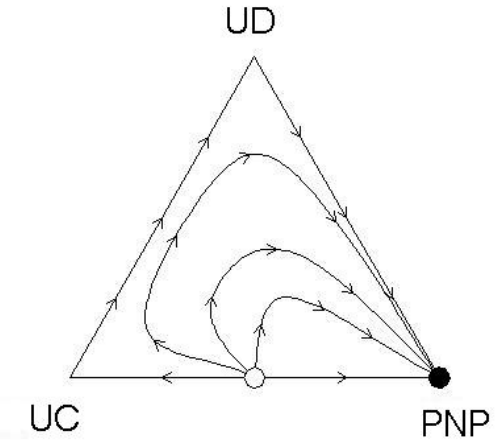

**Figure 4:** *Social dynamics in a population of Cooperators, Defectors and players driven by Symmetric Strong Reciprocity*

***Punishers of (Non-Punishing) Cooperators***

Let us now turn to the dynamics associated with the existence of the 'new' form of strong reciprocity that we label *Punishment of Non-Punishing Cooperators* (*PNP*), as in this case strong reciprocators are willing to incur costs in order to punish cooperators – who unconditionally cooperate but fail to punish defectors, therefore acting as second-order free riders – rather than defectors themselves (for a similar notion, see the seminal paper by Axelrod, 1986). The payoff matrix A, in this case, becomes:

$$A = \begin{array}{c|ccc} & UC & PNP & UD \\ UC & \alpha & \alpha-\varepsilon & \delta \\ PNP & \alpha-\lambda & \alpha & \beta \\ UD & \gamma & \beta & \beta \end{array}$$

Recent evidence from experimental games confirms that cooperative subjects get heavily punished (see e.g. Denant-Boemont et al., 2007; Herrmann et al., 2008; Goette et al., 2010; Abbink et al., 2010)[5]. Also Gächter and Herrmann (2010), in their large-scale experiment with subjects from urban and rural Russia, find a surprisingly high rate of punishment of cooperators: as they correctly point out, "Punishment of cooperators has been largely neglected in previous research on social preferences because it was negligible compared to the punishment of free riders. Our results show that this neglect is not warranted because punishment of cooperators can be very significant in some subject pools". Our dynamic analysis for this case shows that now the pure population stationary state where everyone is a *PNP* is a global attractor in the interior of S and universal cooperation arises (Figure 5). Hence, such a seemingly paradoxical form of sanctioning turns out to be successful in both endogenously enforcing cooperation and being sustainable over time.




[5] Well-known real-life examples of this form of sanctioning include employer's liability for injuries resulting from acts by her employees within the scope of their duties as well as parents' responsibility for harms to others caused by their younger children.

**Figure 5:** *Social dynamics in a population of Cooperators, Defectors and Punishers of (Non-Punishing) Cooperators*

***Hyper-Strong Reciprocators***

We finally illustrate the dynamics in the context in which *Hyper-Strong Negative Reciprocity* (*HSNR*) is present in the population. *HSNR* is a form of strong reciprocity represented by the combination of *SNR* and *PNP*, as *HSNR*s abstain from rewarding other agents (unlike *SPRs*) and incur costs to punish both defectors and cooperators, that is both first-order and second-order free riders. The payoff matrix A associated to this case is:

$$A = \begin{array}{c|ccc} & UC & HSNR & UD \\ UC & \alpha & \alpha-\varepsilon & \delta \\ HSNR & \alpha-\lambda & \alpha & \beta-\lambda \\ UD & \gamma & \beta-\varepsilon & \beta \end{array}$$

When strong reciprocity takes the form of *HSNR*, so that *SRs* display both altruistic punishment of defectors and punishment of non-punishers, in equilibrium either all players become defectors, so that universal defection occurs, or all players become *HSNRs*, so that cooperation flourishes (Figure 6). This case resembles the case of *SNR*, as also in such case we have found that initial conditions turn out to be crucial in order to determine the evolutionary outcome. The key difference, however, is that with *HSNR* the cooperative equilibrium is less fragile than with *SNR*, as it is associated with a monomorphic population rather than with a polymorphic population that can be invaded by defectors[6].

[6] The well-known phenomenon of so called 'collective punishment' provides us with abundant real-life evidence on this variety of strong reciprocity. For example, when something negative happens at school and neither the culprit confesses nor the innocent schoolboys act as informers, the teacher may decide to punish the whole class. As far as *HSNR* is concerned, a significant confirmation of its impact on the enforcement of cooperation in a multi-person prisoner's dilemma is provided by simulation results (Fehr and Fischbacher, 2003).

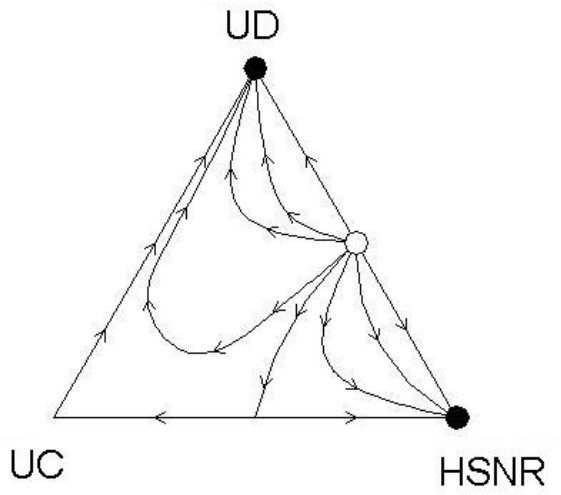



**Figure 6:** *Social dynamics in a population of Cooperators, Defectors and Hyper-Strong Reciprocators*

## 4. Discussion and concluding remarks

In recent years, the idea of strong reciprocity has been gaining more and more credit. The main reason of this seems to lie in the appeal of a notion of endogenous sanctioning based on people's willingness to perform such actions despite the associated monetary costs. Experimental confirmations have generated even more interest towards this behavioral attitude. One of the major results of our evolutionary analysis, however, is that in the three-type population under study, where defectors, cooperators and strong reciprocators initially coexist, if strong reciprocators display *both* altruistic punishment and altruistic rewarding (i.e. *SSR*), in equilibrium all cooperators and punishers perish, so that universal defection eventually prevails. This 'paradox of strong reciprocity', making a behavioral attitude such as *SSR* maladaptive and ineffective as a cooperation enforcement device is due to the 'crowding-out effect' dynamically produced by altruistic rewarding on altruistic punishment: rewarding second-order free riders (that is, cooperators) makes the latter vulnerable and indirectly favors the expansion of first-order free riders (that is, defectors), who effectively exploit cooperators. This makes *SSR* unsustainable and leads to the demise of cooperation. Fehr and Fischbacher (2003) claimed that when cooperation in a population is widespread, altruistic punishers have only a small or no selective disadvantage relative to pure cooperators who abstain from punishing. However, insofar as strong reciprocity means not only costly punishment of defectors but also (symmetrically) costly rewarding of cooperators, our study reveals that – when the proportion of cooperators is extremely large – the cost

disadvantage of strong reciprocators is still relevant due to the fact that rewarding so many cooperators will be costly – though the costs of punishing defectors will be small, in such a circumstance. The second (causally related) problem is that such an increase of cooperators, together with the lack of a large number of strongly reciprocal players around, makes cooperators extremely vulnerable to the 'attack' of defectors, who exploit them and derive relevant advantages from this. This allows them to grow at the expense of cooperators and, eventually, to take over and make the monomorphic pure population equilibrium where all agents defect globally attractive.

Withholding reward to cooperators significantly improves the situation: passing from *SSR* to *SNR* makes strong reciprocity less costly and cooperation sustainable through positive proportions of *SR*s and *UC*s. This is in line with what happens in an indirect reciprocity scenario, where individuals benefit from withholding help. We have also shown that when rewarding does not occur, punishing works better when both cooperators and defectors are sanctioned (under *HSNR*) than when defectors only are sanctioned (*SNR*), as in the latter case, even if costly rewarding does not occur, the locally attractive cooperative equilibria are fragile. However, the adaptive power of strong reciprocity, as well as its capacity to favor the endogenous enforcement of cooperation, are even greater when such behavioral attitude takes the form of *PNP* only, that is when strong reciprocators simply sanction (non-punishing) cooperators and abstain from costly punishing defectors. On the whole, then, our comparative analysis establishes the evolutionary superiority of some varieties of strong reciprocity over others. We have seen that *SNR*, *SPR* and *SSR* perform badly and do not act as effective cooperation enforcement devices, when they have to compete evolutionarily with unconditional cooperation and unconditional defection. By contrast, *PNP* and *HSNR* survive and succeed in enforcing cooperation. Hence, once the inherently plural nature of strong reciprocity is taken into account, it is necessary to specify what is the variety of strong reciprocity one aims at incorporating in a theoretical model based on type heterogeneity, as it would be otherwise impossible to draw unambiguous conclusions about the medium-long run stability of this behavioral attitude.

Why is it the case that, within the same social environment and information scenario, some varieties of strong reciprocity are adaptive while others are not? In a nutshell, our study suggests the following unified answer: in a world in which defectors initially coexist with strong reciprocators and cooperators, the latter can (paradoxically) be an obstacle to the stability of cooperation. The existence of cooperators as prey provides benefits to defectors as predators. Hence, the best way for *SR*s to generate an environment of cooperation and avoid to perish is to try to drive the cooperators to extinction: we show that a strategy by which strongly reciprocal players punish cooperators is highly adaptive both on its own (i.e. *PNP*) and when combined with punishment of first-order free riders (i.e. *HSNR*). On the contrary, a strategy by which *SR*s reward cooperators is highly non-adaptive both on its own (i.e. *SPR*) and, even more so, when combined with punishment of

defectors (i.e. *SSR*). The point is that due to both their being second-order free riders and their failing to reward others, seemingly nice guys such as unconditionally cooperative players are in fact not so nice and deserve the stick, rather than the carrot. These results are in line with recent theoretical work on indirect reciprocity (Ohtsuki et al., 2009) as well as with experimental evidence (Dreber et al., 2008), indicating that subjects who do not punish earn a lot (they are the 'winners'), while punishers end up with low payoff levels (they are the 'losers'). Hence, punishment of cooperators becomes itself socially beneficial and, therefore, 'altruistic', while rewarding cooperators is socially harmful and can be viewed a 'antisocial'.

Our findings suggest that cooperation can emerge due to the crucial role played by strong reciprocity but also that, in societies with sizeable shares of unconditional cooperators, strong reciprocity can be successful insofar as it takes the form of 'punishment of cooperators'. Such an evolutionary account of cooperation is based on an individual selection framework and is compatible with the presence, in the population, of cooperative 'good men' who, by doing nothing, risk to favor the 'triumph of evil' (as the poet Burke famously put it): unlike theories of cooperation based on altruistic punishment of defectors only, this explanation takes into account the 'dark' side of (seemingly) other-regarding behavior and sheds light on the potential role of a plural behavioral attitude such as strong reciprocity in effectively dealing with it.

## Mathematical appendix

We analyze dynamics (2) by using Bomze's (1983) classification for replicator equations. In order to present social dynamics for all the five varieties of strong reciprocity we focus on, we have to consider five distinct material payoff matrices, in correspondence with the five three-type populations under study, on the basis of the material consequences from the two-stage *PDCS* game conveyed by Tables 1 and 2. All the five cases illustrated in the main text of the paper are analyzed on the basis of the propositions we state here below. In order to use Bomze's classification, we need to re-write the payoff matrix (1) in the following form:

$$B=\begin{pmatrix} 0 & 0 & 0 \\ a & b & c \\ d & e & f \end{pmatrix}=\begin{pmatrix} 0 & 0 & 0 \\ a_{21}-a_{11} & a_{22}-a_{12} & a_{23}-a_{13} \\ a_{31}-a_{11} & a_{32}-a_{12} & a_{33}-a_{13} \end{pmatrix} \quad (3)$$

with the first row made of zeros[7]. Dynamics (2) is equivalent (Hofbauer 1981) to the Lotka-Volterra system:

$$\begin{aligned} \dot{x} &= x(a + bx + cy) \\ \dot{y} &= y\ (d + ex + fy) \end{aligned} \tag{4}$$

by the coordinate transformation:

$$x_1 = \frac{1}{1+x+y}, \quad x_2 = \frac{x}{1+x+y}, \quad x_3 = \frac{y}{1+x+y} \tag{5}$$

This coordinate change is sometimes used in the analysis below. Furthermore, we make use of the same terminology used in Bomze (1983). By an eigenvalue EV of a stationary state we shall understand an eigenvalue of the linearization matrix around that stationary state. The term *EV in direction of the vector V* means that *V* is an eigenvector corresponding to that *EV*. *IntS* is the set $\{x \in S : x_i > 0,\ i = 1, 2, 3\}$ in which all behavioral types are present in the population. An *edge* of *S* consists of all population states in which a given (fixed) strategy is not adopted; we shall denote $K_{ij}$ the edge joining $e_i$ with $e_j$, where $e_1 = (1,0,0)$, $e_2 = (0,1,0)$, $e_3 = (0,0,1)$ are the vectors of the canonical basis; $e_1$, $e_2$ and $e_3$ represent the states in which in the population there are only *UC*s, *SR*s and *UD*s, respectively. Thus e.g. $K_{12}$ is the edge where only types *UC* and SR are present in the population. Note that, by (5), $K_{12}$ corresponds to the positive semi-axis $y = 0$ of the plane (x,y) and $K_{13}$ corresponds to the positive semi-axis $x = 0$.

The stability properties of the vertices $e_1$, $e_2$ and $e_3$ (indicated, respectively, by *UC*, *SR* and *UD* in Figures 2-6) are analyzed in the following proposition[8]. For simplicity, the propositions in Bomze (1983) will be indicated as B# (so, e.g., B4 is Proposition 4 of Bomze's paper).

**Proposition 1** *The eigenvalue structure of the stationary states $e_i$ is the following:*

*(1) $e_1$ has one eigenvalue with the sign of $a$ in direction of $K_{12}$ and one eigenvalue with the sign of $d$ in direction of $K_{13}$.*

*(2) $e_2$ has one eigenvalue with the sign of $-b$ in direction of $K_{12}$ and one eigenvalue with the sign of $e-b$ in direction of $K_{23}$.*

---

[7] It is a well-known result that dynamics (2) does not change if an arbitrary constant is added to each column of *A* (see e.g. Hofbauer and Sigmund, 1988; p. 126).

[8] All the eigenvalues of the stationary stateson the edges of *S* are real (see Bomze, 1983).

*(3)* $e_3$ *has one eigenvalue with the sign of* $-f$ *in direction of* $K_{13}$ *and one eigenvalue with the sign of* $c-f$ *in direction of* $K_{23}$.

**Proof.** See B1.

The following proposition concerns the stationary states on the edges of *S*.

**Proposition 2** *(1)* $K_{12}$ *is pointwise fixed*[9] *if and only if (iff)* $a=b=0$. *There is a unique the stationary state in the interior of* $K_{12}$ *iff* $ab<0$. *In the remaining cases, there are no the stationary states in it. The eigenvalues of the unique the stationary state (when existing) have the sign of* $-a$ *in direction of* $K_{12}$ *and of* $(bd-ae)/b$ *in direction of the interior of S.*

*(2) There are not the stationary states in* $K_{13}$.

*(3) There is a unique the stationary state in the interior of* $K_{23}$ *iff* $(e-b)(f-c)<0$. *In the remaining cases, there are not the stationary states in* $K_{23}$. *The eigenvalues of the unique the stationary state in the interior of* $K_{23}$ *have the sign of* $(e-b)(f-c)/(e-b+c-f)$ *in direction of* $K_{23}$ *and of* $(bf-ce)/(e-b+c-f)$ *in direction of the interior of S.*

**Proof.** Apply B2 and B5.

The remaining proposition concerns the stationary states in the interior of *S*, where all behavioral types coexist.

**Proposition 3** *There is a unique the stationary state in IntS iff the expressions:*

$$bf-ce \qquad ae-bd \qquad cd-af \tag{6}$$

*have all the same sign and are not equal to zero.*

*If they are all zero, then there is a pointwise fixed line* $G=\{(x,y): a+bx+cy=d+ex+fy=0\}$ *in IntS (if the intersection between G and the positive quadrant of the plane* $(x,y)$ *is not empty). In the remaining cases, there are not stationary states in IntS.*

**Proof.** Apply B6.

**Strong Negative Reciprocity (*SNR*)**

---

[9] The term pointwise fixed means that all the points of such an edge are stationary states.

In such a case, the payoff matrix *B* (see (3)) becomes:

$$B=\begin{pmatrix} 0 & 0 & 0 \\ a & b & c \\ d & e & f \end{pmatrix}=\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \beta-\delta-\lambda \\ \gamma-\alpha & \beta-\alpha-\varepsilon & \beta-\delta \end{pmatrix}$$

In this case, by Propositions 1-3, we have that $e_3$ (where all players are *UD*s) is always locally attractive and the edge $K_{12}$ (where *UD*s are not present) is always pointwise fixed. Furthermore, there is always an isolated stationary state in the edge $K_{23}$ which has two positive eigenvalues if $\lambda > \beta - \delta$, one positive eigenvalue (in direction of $K_{23}$) and one negative eigenvalue in direction of *IntS* if $\lambda < \beta - \delta$. In the interior of *S* there are not isolated stationary states while there exists a pointwise fixed line:

$$y = -\frac{\gamma-\alpha}{\beta-\delta} + \frac{\alpha+\varepsilon-\beta}{\beta-\delta}x \tag{7}$$

iff $\lambda = \beta - \delta$. Looking at all possible dynamic regimes showed in Bomze (1983), we can classify the dynamics in a *UC-SNR-UD* population. In particular, if $\lambda < \beta - \delta$ we have the phase portrait number 28 in Bomze's (1983) classification (Bpp#, hereafter); if $\lambda = \beta - \delta$ we obtain Bpp3 and if $\lambda > \beta - \delta$ we obtain Bpp23. For simplicity, in the paper we show only the dynamics related to the latter case (for the other two cases, qualitative dynamics is very similar). Consequently, in a *UC-SNR-UD* population, a bi-stable dynamics always emerges: the trajectories in the interior of S converge either to the stationary state $e_3$, where all players are *UDs*, or to the edge $K_{12}$, where *UDs* disappear and *UCs* and *SNRs* coexist (see Figure 2).

The following proposition leads to an intuitive result concerning the variations of the basins of attraction of $e_3$ and $K_{12}$ when the parameter $\varepsilon$ (capturing defectors' cost of being punished by strong reciprocators) varies.

**Proposition 4** *As $\varepsilon$ increases, the basin of attraction of the stationary state $e_3$ gets smaller and, as a consequence, the basin of attraction of $K_{12}$ gets larger.*

**Proof.** In the case of *SNR*, if we write the dynamics in the coordinates (*x*,*y*) (see (5)), we obtain:

$$\begin{aligned} \dot{x} &= x(\beta-\lambda-\delta)y \\ \dot{y} &= y[\gamma-\alpha+(\beta-\varepsilon-\alpha)x+(\beta-\delta)y] \end{aligned} \tag{8}$$

Therefore, the slope of the trajectories is given by the following expression:

$$\frac{dy}{dx} = \frac{\dot{y}}{\dot{x}} = -\frac{\alpha + \varepsilon - \beta}{\beta - \lambda - \delta} + \frac{\gamma - \alpha + (\beta - \delta)y}{(\beta - \lambda - \delta)x} \tag{9}$$

Let us first consider the case $\lambda < \beta - \delta$; the basins of attraction of $e_3$ and $K_{12}$ are separated by the unique trajectories converging to the stationary state on the edge $K_{23}$ (see Bpp28 in Bomze, 1983). From (9) we get that passing from $\varepsilon_1$ to $\varepsilon_2$, with $\varepsilon_1 < \varepsilon_2$, we obtain:

$$\left.\frac{dy}{dx}\right|_{\varepsilon = \varepsilon_1} > \left.\frac{dy}{dx}\right|_{\varepsilon = \varepsilon_2}$$

This implies that the trajectory $\Gamma$ converging to the stationary state in $K_{23}$ for $\varepsilon = \varepsilon_1$ gets crossed top-down from the trajectories of (8) for $\varepsilon = \varepsilon_2$. This implies that the basin of $e_3$ gets smaller and the basin of $K_{12}$ gets larger, passing from $\varepsilon_1$ to $\varepsilon_2$.

Let us now turn to the case $\lambda = \beta - \delta$ (see Bpp3 in Bomze, 1983); the basins of attraction of $e_3$ and $K_{12}$ are separated by the line (7). Since such line shifts upward (in the positive quadrant of the plane ($x$, $y$)) as $\varepsilon$ grows, we proved the proposition for such a dynamic regime as well.

Let us now turn to the case $\lambda > \beta - \delta$; in this case, there is only one trajectory starting from the stationary state in $K_{23}$ (which is a repulsive) tangent to the edge $K_{12}$; the part $\Phi$ of this trajectory that goes from the stationary state in $K_{23}$ to the edge $K_{12}$ separates the basins of attraction of $e_3$ and of $K_{12}$. Since for $\lambda > \beta - \delta$ we have:

$$\left.\frac{dy}{dx}\right|_{\varepsilon = \varepsilon_1} < \left.\frac{dy}{dx}\right|_{\varepsilon = \varepsilon_2}$$

then the trajectory $\Phi$ for $\varepsilon = \varepsilon_1$ gets crossed top-down from the trajectories of (8) for $\varepsilon = \varepsilon_2$. This implies that the basin of $e_3$ gets smaller and the basin of $K_{12}$ gets larger in passing from $\varepsilon_1$ to $\varepsilon_2$.

**Strong Positive Reciprocity (*SPR*)**

In order to avoid a lengthy presentation of our calculations, we omit to write the payoff matrices B for this case and for the following ones (the payoff matrices A are given in the main text). The procedure that allows us to set them up should be clear now. The dynamics in a *UC-SPR-UD* population is characterized by the following properties. There are not stationary states in the edges $K_{ij}$ of S and all the vertices $e_i$ are saddle points. Furthermore, there always exists one stationary state in the interior of S; by applying Corollary 7 of B6 in Bomze (1983), it is easy to show that such a point is a source. Thus, in the case of a *UC-SPR-UD*

population, we have the regime shown in Figure 3: all trajectories approach the boundary of *S* turning clockwise.

**Symmetric Strong Reciprocity (*SSR*)**

We have that $e_3$ (where all players are *UD*s) is always locally attractive. There are not stationary states in the edges $K_{12}$ and $K_{13}$; there is always an isolated stationary state in the edge $K_{23}$ which has one positive eigenvalue in direction of $K_{23}$. In the interior of *S*, pointwise fixed lines do not exist (being $ae-bd>0$ always) and one stationary state exists iff $(\beta-\delta-\lambda)(\alpha+\varepsilon+\eta-\beta)>\eta(\beta-\delta)$; in the other cases, no stationary state exists in the interior of *S*. Consequently, by Bomze's classification, we obtain two dynamic regimes. In particular, when there exists a stationary state in the interior of *S*, the dynamic regime is Bpp15 in Bomze (1983); if it does not exist, the dynamic regime is the one showed in Figure 3. In both dynamic regimes, the stationary state $e_3$ is globally attractive, so that we obtain the result illustrated in the paper.

**Punishment of Non-Punishing Cooperators (*PNP*)**

In this context, the stationary state $e_2$ (where all players are *PNPs*) is always attractive while the other vertices are saddles; there are not stationary states in the edges $K_{13}$ and $K_{23}$ and there is always one stationary state in $K_{12}$ which has a positive eigenvalue in direction of $K_{12}$. Furthermore, there is a unique stationary state in the interior of S iff $\lambda(\beta+\varepsilon-\alpha)<\varepsilon(\alpha-\gamma)$; in such a case, the stationary state in $K_{12}$ is a saddle. If $\lambda(\beta+\varepsilon-\alpha)\geq\varepsilon(\alpha-\gamma)$, there are not stationary states in the interior of S and the stationary state in $K_{12}$ is repulsive. If the interior stationary state exists, the associated dynamic regime is Bpp15; otherwise, the dynamic regime is the one showed in Figure 5. In both regimes, the stationary state $e_2$ is a global attractor.

**Hyper-Strong Negative Reciprocity (*HSNR*)**

Here, the vertices $e_2$ (where all players are *HSNR*s) and $e_3$ (where all players are *UD*s) are always attractive while $e_1$ is always a saddle. There are not stationary states in $K_{13}$ and there is one stationary state in $K_{12}$ and one in $K_{23}$; both these stationary states have a positive eigenvalue in direction of $K_{12}$ and $K_{23}$, respectively. The dynamic regimes, in the case of a *UC-HSNR-UD* population, depend on the sign of the following expressions (see Proposition 3):

$$bf-ce=\varepsilon(\beta-\delta)+(\alpha-\beta)(\beta-\delta-\lambda)$$
$$ae-bd=\lambda(\alpha-\beta)+\varepsilon(\alpha-\gamma)$$

$$cd - af = (\gamma - \alpha)(\beta - \delta - \lambda) + \lambda(\beta - \delta)$$

In particular, we have the following sub-cases:

1) If $bf - ce > 0$, $ae - bd > 0$, $cd - af > 0$, then there exists one stationary state in the interior of S which is repulsive (see B6 and Corollary 7) and the stationary states in $K_{12}$ and in $K_{23}$ are saddles. The corresponding dynamic regime is Bpp9.

2) If $bf - ce = ae - bd = cd - af = 0$, then there exists one pointwise fixed line in the interior of S joining the stationary states in $K_{12}$ and in $K_{23}$. The corresponding dynamic regime is Bpp5.

3) $bf - ce < 0$, $ae - bd < 0$, $cd - af < 0$, then there exists one stationary state in the interior of S which is a saddle (see B6 and Corollary 7) and the stationary states in $K_{12}$ and in $K_{23}$ are repulsive. The corresponding dynamic regime is Bpp10.

4) If $bf - ce > 0$, $ae - bd \leq 0$, $\forall\, cd - af$, then there are not stationary states in the interior of S, the stationary state in $K_{12}$ is repulsive and that in $K_{23}$ is a saddle point. The corresponding dynamic regime is Bpp37.

5) If $bf - ce \leq 0$, $ae - bd > 0$, $\forall\, cd - af$, then there are not stationary states in the interior of S, the stationary state in $K_{12}$ is a saddle point and that in $K_{23}$ is repulsive. The corresponding dynamic regime is Bpp38.

A representative scenario, with *HSNR*, is the one showed in Figure 6, corresponding to the phase portrait Bpp38.

## References


Abbink, K., Irlenbusch, B., Renner, E., 2000, The moonlighting game: an experimental study on reciprocity and retribution, *Journal of Economic Behavior and Organization*, 2, 42, 265-77.

Abbink, K., Brandts, J., Herrmann, B., Orzen, H., 2010, Intergroup conflict and intra-group punishment in an experimental contest game, *American Economic Review*, 100, 420-447.

Axelrod, R., 1986, An evolutionary approach to norms, *American Political Science Review*, 80, 4, 1095-1111.

Axelrod, R.., 1984, The Evolution of Cooperation, Basic Books.

Bomze, I., 1983, Lotka-Volterra equations and replicator dynamics: a two-dimensional classification, *Biological Cybernetics*, 48, 201-11.

Camera, G., Casari, M., 2009, Cooperation among strangers under the shadow of the future. *American Economic Review*, 99, 3, 979-1005.

Carpenter, J.P., Matthews, P.H., Okmomboli, O., 2004, Why punish? Social reciprocity and the enforcement of prosocial norms, *Journal of Evolutionary Economics*, 14, 4, 407-429.

Denant-Boemont, L., Masclet, D., Noussair, C.N., 2007, Punishment, counterpunishment and sanction enforcement in a social dilemma experiment, *Economic Theory*, 33, 145-167.

Dreber, A., Rand, D.G., Fudenberg, D., Nowak, M.A., 2008, Winners don't punish, *Nature*, 452, 348-351.

Eshel, I., Samuelson, L., Shaked, A., 1998, Altruists, egoists, and hooligans in a local interaction model, *American Economic Review*, 88, 157-179.

Fehr, E., Fischbacher, U., 2003, The nature of human altruism, *Nature* 425, 785-791.

Fehr, E., Fischbacher, U., 2005, The Economics of Strong Reciprocity, in Gintis H., Bowles S., Boyd R. and Fehr E. (eds), Moral Sentiments and Material Interests. The Foundations of Cooperation in Economic Life, Cambridge (Mass.) and London, MIT Press, 151-91.

Fehr, E., Fischbacher, U., Gächter, S., 2002, Strong reciprocity, human cooperation and the enforcement of social norms, *Nature*, 13, 1-25.

Fehr, E., Gächter, S., 2000, Cooperation and punishment, *American Economic Review*, 90, 4, 980-94.

Fehr, E., Henrich, J., 2003, Is Strong Reciprocity a Maladaptation? in The Genetic and Cultural Evolution of Cooperation (ed Hammerstein, P.), MIT Press.

Fischbacher, U., Gächter, S., 2010, Social preferences, beliefs, and the dynamics of free riding in public good experiments, *American Economic Review*, 100 (1), 541-556.

Fowler, J.H., 2005, Altruistic punishment and the origin of cooperation, *Proceedings of the National Academy of Sciences* 102, 19, 7047-7049.

Fudenberg, D., Maskin, E., 1986, The folk theorem in repeated games with discounting or with incomplete information, *Econometrica*, 50, 533-554.

Gächter, S., Herrmann, B., 2010, The limits of self-governance when cooperators get punished: experimental evidence from urban and rural Russia, *European Economic Review*, 55 (2), 193-210.

Gintis, H., Bowles, S., Boyd, R., Fehr, E., 2005. Moral Sentiments and Material Interests: Origins, Evidence, and Consequences, in Moral Sentiments and Material Interests. The Foundations of Cooperation in Economic Life (eds Gintis, H., Bowles, S., Boyd, R., Fehr, E.), MIT Press, 3-39.

Goette, L., Huffman, D., Meier, S., Sutter, M., 2010, Group membership, competition, and altruistic versus

antisocial punishment: evidence from randomly assigned army groups, IZA Discussion Paper N. 5189.

Hamilton, W.D., 1964, The genetical evolution of social behavior, *Journal of Theoretical Biology,* 7, 1-16.

Herrmann, B., Thoeni, C., Gächter, S., 2008, Antisocial punishment across societies, *Science,* 319, 1362-1367.

Hofbauer J., 1981, On the Occurrence of Limit Cycles in the Volterra-Lotka Equation, Nonlinear Analysis. Theory, Methods and Applications, 5, 1003-1007.

Hofbauer, J., Sigmund, K., 1988, The Theory of Evolution and Dynamical Systems, Cambridge, Cambridge University Press.

Nowak, M.A., Sigmund, K., 1998, Evolution of indirect reciprocity by image scoring, *Nature,* 393, 573-577.

Offerman, T., 2002, Hurting hurts more than helping helps, *European Economic Review,* 46, 1423-37.

Ohtsuki, H., Iwasa, Y. & Nowak, M.A., 2009, Indirect reciprocity provides only a narrow margin of efficiency for costly punishment, *Nature,* 457, 79-82.

Ones, U., Putterman, L., 2007, The ecology of collective action: a public goods and sanctions experiment with controlled group formation, *Journal of Economic Behavior and Organization,* 62, 4, 495-521.

Panchanathan, K., Boyd, R., 2004, Indirect reciprocity can stabilize cooperation without the second-order free rider problem, *Nature* 432, 499-502.

Rand, D.G., Armao IV, J.J., Nakamaru, M., Ohtsuki, H., 2010, Anti-social punishment can prevent the co-evolution of punishment and cooperation, *Journal of Theoretical Biology*, 265, 624-632.

Reuben, E., van Winden, F., 2010, Fairness perceptions and prosocial emotions in the power to take, *Journal of Economic Psychology,* 31, 908-922.

Sethi, R., Somanathan, E., 1996, The evolution of social norms in common property resource use, *American Economic Review,* 86, 4, 766-788.

Sutter, M., Haigner, S., Kocher, M.G., 2010, Choosing the carrot or the stick? Endogenous institutional choice in social dilemma situations, *Review of Economic Studies,* 77, 4, 1540-1566.

Taylor P., Yonker L., 1978, Evolutionary stable strategies and game dynamics, *Mathematical Biosciences,* 40, 145-56.

Weibull, J.W., 1995, Evolutionary Game Theory, Cambridge (Ma.), MIT Press.

**NOTE DI LAVORO DELLA FONDAZIONE ENI ENRICO MATTEI**

**Fondazione Eni Enrico Mattei Working Paper Series**

**Our Note di Lavoro are available on the Internet at the following addresses:**

http://www.feem.it/getpage.aspx?id=73&sez=Publications&padre=20&tab=1
http://papers.ssrn.com/sol3/JELJOUR_Results.cfm?form_name=journalbrowse&journal_id=266659
http://ideas.repec.org/s/fem/femwpa.html
http://www.econis.eu/LNG=EN/FAM?PPN=505954494
http://ageconsearch.umn.edu/handle/35978
http://www.bepress.com/feem/

**NOTE DI LAVORO PUBLISHED IN 2011**

| | | |
|---|---|---|
| SD | 1.2011 | Anna Alberini, Will Gans and Daniel Velez-Lopez: Residential Consumption of Gas and Electricity in the U.S.: The Role of Prices and Income |
| SD | 2.2011 | Alexander Golub, Daiju Narita and Matthias G.W. Schmidt: Uncertainty in Integrated Assessment Models of Climate Change: Alternative Analytical Approaches |
| SD | 3.2010 | Reyer Gerlagh and Nicole A. Mathys: Energy Abundance, Trade and Industry Location |
| SD | 4.2010 | Melania Michetti and Renato Nunes Rosa: Afforestation and Timber Management Compliance Strategies in Climate Policy. A Computable General Equilibrium Analysis |
| SD | 5.2011 | Hassan Benchekroun and Amrita Ray Chaudhuri: "The Voracity Effect" and Climate Change: The Impact of Clean Technologies |
| IM | 6.2011 | Sergio Mariotti, Marco Mutinelli, Marcella Nicolini and Lucia Piscitello: Productivity Spillovers from Foreign MNEs on Domestic Manufacturing Firms: Is Co-location Always a Plus? |
| GC | 7.2011 | Marco Percoco: The Fight Against Geography: Malaria and Economic Development in Italian Regions |
| GC | 8.2011 | Bin Dong and Benno Torgler: Democracy, Property Rights, Income Equality, and Corruption |
| GC | 9.2011 | Bin Dong and Benno Torgler: Corruption and Social Interaction: Evidence from China |
| SD | 10.2011 | Elisa Lanzi, Elena Verdolini and Ivan Haščič: Efficiency Improving Fossil Fuel Technologies for Electricity Generation: Data Selection and Trends |
| SD | 11.2011 | Stergios Athanassoglou: Efficient Random Assignment under a Combination of Ordinal and Cardinal Information on Preferences |
| SD | 12.2011 | Robin Cross, Andrew J. Plantinga and Robert N. Stavins: The Value of Terroir: Hedonic Estimation of Vineyard Sale Prices |
| SD | 13.2011 | Charles F. Mason and Andrew J. Plantinga: Contracting for Impure Public Goods: Carbon Offsets and Additionality |
| SD | 14.2011 | Alain Ayong Le Kama, Aude Pommeret and Fabien Prieur: Optimal Emission Policy under the Risk of Irreversible Pollution |
| SD | 15.2011 | Philippe Quirion, Julie Rozenberg, Olivier Sassi and Adrien Vogt-Schilb: How CO2 Capture and Storage Can Mitigate Carbon Leakage |
| SD | 16.2011 | Carlo Carraro and Emanuele Massetti: Energy and Climate Change in China |
| SD | 17.2011 | ZhongXiang Zhang: Effective Environmental Protection in the Context of Government Decentralization |
| SD | 18.2011 | Stergios Athanassoglou and Anastasios Xepapadeas: Pollution Control with Uncertain Stock Dynamics: When, and How, to be Precautious |
| SD | 19.2011 | Jūratė Jaraitė and Corrado Di Maria: Efficiency, Productivity and Environmental Policy: A Case Study of Power Generation in the EU |
| SD | 20.2011 | Giulio Cainelli, Massimiliano Mozzanti and Sandro Montresor: Environmental Innovations, Local Networks and Internationalization |
| SD | 21.2011 | Gérard Mondello: Hazardous Activities and Civil Strict Liability: The Regulator's Dilemma |
| SD | 22.2011 | Haiyan Xu and ZhongXiang Zhang: A Trend Deduction Model of Fluctuating Oil Prices |
| SD | 23.2011 | Athanasios Lapatinas, Anastasia Litina and Eftichios S. Sartzetakis: Corruption and Environmental Policy: An Alternative Perspective |
| SD | 24.2011 | Emanuele Massetti: A Tale of Two Countries:Emissions Scenarios for China and India |
| SD | 25.2011 | Xavier Pautrel: Abatement Technology and the Environment-Growth Nexus with Education |
| SD | 26.2011 | Dionysis Latinopoulos and Eftichios Sartzetakis: Optimal Exploitation of Groundwater and the Potential for a Tradable Permit System in Irrigated Agriculture |
| SD | 27.2011 | Benno Torgler and Marco Piatti. A Century of *American Economic Review* |
| SD | 28.2011 | Stergios Athanassoglou, Glenn Sheriff, Tobias Siegfried and Woonghee Tim Huh: Optimal Mechanisms for Heterogeneous Multi-cell Aquifers |
| SD | 29.2011 | Libo Wu, Jing Li and ZhongXiang Zhang: Inflationary Effect of Oil-Price Shocks in an Imperfect Market: A Partial Transmission Input-output Analysis |
| SD | 30.2011 | Junko Mochizuki and ZhongXiang Zhang: Environmental Security and its Implications for China's Foreign Relations |
| SD | 31.2011 | Teng Fei, He Jiankun, Pan Xunzhang and Zhang Chi: How to Measure Carbon Equity: Carbon Gini Index Based on Historical Cumulative Emission Per Capita |
| SD | 32.2011 | Dirk Rübbelke and Pia Weiss: Environmental Regulations, Market Structure and Technological Progress in Renewable Energy Technology — A Panel Data Study on Wind Turbines |

| | | |
|---|---|---|
| SD | 33.2011 | Nicola Doni and Giorgio Ricchiuti: Market Equilibrium in the Presence of Green Consumers and Responsible Firms: a Comparative Statics Analysis |
| SD | 34.2011 | Gérard Mondello: Civil Liability, Safety and Nuclear Parks: Is Concentrated Management Better? |
| SD | 35.2011 | Walid Marrouch and Amrita Ray Chaudhuri: International Environmental Agreements in the Presence of Adaptation |
| ERM | 36.2011 | Will Gans, Anna Alberini and Alberto Longo: Smart Meter Devices and The Effect of Feedback on Residential Electricity Consumption: Evidence from a Natural Experiment in Northern Ireland |
| ERM | 37.2011 | William K. Jaeger and Thorsten M. Egelkraut: Biofuel Economics in a Setting of Multiple Objectives & Unintended Consequences |
| CCSD | 38.2011 | Kyriaki Remoundou, Fikret Adaman, Phoebe Koundouri and Paulo A.L.D. Nunes: Are Preferences for Environmental Quality Sensitive to Financial Funding Schemes? Evidence from a Marine Restoration Programme in the Black Sea |
| CCSD | 39.2011 | Andrea Ghermanti and Paulo A.L.D. Nunes: A Global Map of Costal Recreation Values: Results From a Spatially Explicit Based Meta-Analysis |
| CCSD | 40.2011 | Andries Richter, Anne Maria Eikeset, Daan van Soest, and Nils Chr. Stenseth: Towards the Optimal Management of the Northeast Arctic Cod Fishery |
| CCSD | 41.2011 | Florian M. Biermann: A Measure to Compare Matchings in Marriage Markets |
| CCSD | 42.2011 | Timo Hiller: Alliance Formation and Coercion in Networks |
| CCSD | 43.2011 | Sunghoon Hong: Strategic Network Interdiction |
| CCSD | 44.2011 | Arnold Polanski and Emiliya A. Lazarova: Dynamic Multilateral Markets |
| CCSD | 45.2011 | Marco Mantovani, Georg Kirchsteiger, Ana Mauleon and Vincent Vannetelbosch: Myopic or Farsighted? An Experiment on Network Formation |
| CCSD | 46.2011 | Rémy Oddou: The Effect of Spillovers and Congestion on the Segregative Properties of Endogenous Jurisdiction Structure Formation |
| CCSD | 47.2011 | Emanuele Massetti and Elena Claire Ricci: Super-Grids and Concentrated Solar Power: A Scenario Analysis with the WITCH Model |
| ERM | 48.2011 | Matthias Kalkuhl, Ottmar Edenhofer and Kai Lessmann: Renewable Energy Subsidies: Second-Best Policy or Fatal Aberration for Mitigation? |
| CCSD | 49.2011 | ZhongXiang Zhang: Breaking the Impasse in International Climate Negotiations: A New Direction for Currently Flawed Negotiations and a Roadmap for China to 2050 |
| CCSD | 50.2011 | Emanuele Massetti and Robert Mendelsohn: Estimating Ricardian Models With Panel Data |
| CCSD | 51.2011 | Y. Hossein Farzin and Kelly A. Grogan: Socioeconomic Factors and Water Quality in California |
| CCSD | 52.2011 | Dinko Dimitrov and Shao Chin Sung: Size Monotonicity and Stability of the Core in Hedonic Games |
| ES | 53.2011 | Giovanni Mastrobuoni and Paolo Pinotti: Migration Restrictions and Criminal Behavior: Evidence from a Natural Experiment |
| ERM | 54.2011 | Alessandro Cologni and Matteo Manera: On the Economic Determinants of Oil Production. Theoretical Analysis and Empirical Evidence for Small Exporting Countries |
| ERM | 55.2011 | Alessandro Cologni and Matteo Manera: Exogenous Oil Shocks, Fiscal Policy and Sector Reallocations in Oil Producing Countries |
| ERM | 56.2011 | Morgan Bazilian, Patrick Nussbaumer, Giorgio Gualberti, Erik Haites, Michael Levi, Judy Siegel, Daniel M. Kammen and Joergen Fenhann: Informing the Financing of Universal Energy Access: An Assessment of Current Flows |
| CCSD | 57.2011 | Carlo Orecchia and Maria Elisabetta Tessitore: Economic Growth and the Environment with Clean and Dirty Consumption |
| ERM | 58.2011 | Wan-Jung Chou, Andrea Bigano, Alistair Hunt, Stephane La Branche, Anil Markandya and Roberta Pierfederici: Households' WTP for the Reliability of Gas Supply |
| ES | 59.2011 | Maria Comune, Alireza Naghavi and Giovanni Prarolo: Intellectual Property Rights and South-North Formation of Global Innovation Networks |
| ES | 60.2011 | Alireza Naghavi and Chiara Strozzi: Intellectual Property Rights, Migration, and Diaspora |
| CCSD | 61.2011 | Massimo Tavoni, Shoibal Chakravarty and Robert Socolow: Safe vs. Fair: A Formidable Trade-off in Tackling Climate Change |
| CCSD | 62.2011 | Donatella Baiardi, Matteo Manera and Mario Menegatti: Consumption and Precautionary Saving: An Empirical Analysis under Both Financial and Environmental Risks |
| ERM | 63.2011 | Caterina Gennaioli and Massimo Tavoni: Clean or "Dirty" Energy: Evidence on a Renewable Energy Resource Curse |
| ES | 64.2011 | Angelo Antoci and Luca Zarri: Punish and Perish? |