The economic analysis of a Q-learning model of Cooperation with punishment.

Solferino, Nazaria and Solferino, Viviana and Taurino, Serena Fiona

University of Tor Vergata in Rome, University of Calabria, University of Tor Vergata in Rome

11 September 2015

The Economics analysis of a Q-learning model of Cooperation with Punishment

Nazaria Solferino

Economics Department, University of Rome "Tor Vergata"

Viviana Solferino

Mathematics and Computer Science Department, University of Calabria

Serena F. Taurino

Economics Department, University of Rome "Tor Vergata"

Abstract

A Q-learning model is devised in order to see whether individuals can "learn" how to cooperate, when a virtuous system of punishment and reinforcement is adopted. The paper shows that if it is possible to free-ride and not being adequately punished, there will always be an incentive to deviate from the cooperation.Conversely, even if the others did not cooperate, it is still possible to have someone who cooperates when individuals are pushed by strong intrinsic motivation. Cooperation can be a learning process. It is possible to trigger a learning process that leads individuals to be equally cooperative. This happens much more easily, the more responsible individuals are. It also depends on proper punishment.

Keywords: Cooperation, Punishment, Q-learning models.

Jel Numbers: C70, C72, D62, C27

1 Introduction

It is generally believed that public goods can be produced only in the presence of repeated interactions (which allow reciprocation, reputation effects and punishment) or relatedness. In a game context with the minimum threshold for public goods, a minimum amount of contributions from the participants has to be collected for the providing the public good to occur. Nevertheless, the production of public goods by the contribution of individual volunteers is a social dilemma: an individual can benefit from the public good produced by the contributions of others even if not volunteering.

To this aim, the occurrence and maintenance of cooperative behaviors in public goods systems have attracted great research attention across multiple disciplines. Mechanisms that allow the rise and maintenance of cooperation have been analysed by a conspicuous literature, also when in the presence of defectors (Dawes, 1980; Hardin, 1968, Kagel and Roth, 1997). Boyd and Richerson (1998) describe how recurrent interactions among individuals in potentially cooperative situations are likely to evolve into a stable reciprocal cooperation. But the increase in group size and number of potential defectors make conditions extremely restrictive. Also, numerical simulations of the infinitely iterated stochastic games (Hauert and Schuster; 1998) give evidence to the fact that stable cooperative solutions are strong strategies. They are barely influenced by memory size and different values of the temptation to defect. Corresponding results are into the analysis - performed by Schuster and Sigmund (1983) - of several evolutionary models in distinct biological fields.

Other studies (Boyd and Richerson,1992; Fehr and Gachter, 2000) show how the promotion of cooperation, as well as defector punishment, can prevent the end of cooperation with defection going as prevalent strategy. Additionally, when voluntary participation and altruistic punishment of each defector work together, they support the emergence and the stabilization of cooperation.

Theoretical (Fowler, 2005; Hauert et al., 2007; Nakamaru and Dieckmann,2009; Sigmund et al.,2010; Sasaki et al.,2012; Brandt et al., 2006, Hauert et al., 2007, 2008) and experimental papers (Egas and Riedl, 2008; Fehr and Gachter, 2002) have showed these working, under the hypothesis of perfect information about players' strategies. Among them, Boyd

and Richerson (1992) state that the combination of punishment (both to a defector and irresponsible institutions or officers) and ethical strategies is progressively stable.

An experimental work from Fehr and Gachter (2000) explains how cooperation prospers when it is possible to have altruistic punishment and interrupts if its continuity is broken. Another work by Brandt et al. (2006) grounds a bi-stable result onto a microeconomic model. They show how you can have evolutionary dynamics going to a Nash equilibrium with punishment, non-punishment strategy, as well as to an oscillating state without punishers. Punishment of defector is the base for the beginning and the constitution of cooperative behaviour as for the work of Hauert et al.(2007). Also, they highlight as the free and choral choice by all players of punishing non-cooperators is necessary to have such mechanism to work. Another contribution from Nakamaru and Dieckmann (2009) points out like runaway selection can emerge from punishment and cooperation, leading to increased collaboration. They also show how such increase is stronger the lower the cost of punishment.

Sigmund et al.(2010) find that pool-punishment is more efficient than peer-punishment in preventing from second-order free-riders. It is so, as this type of free-riders are active even if every single individual is contributing to the common good. Sasaki et al.(2012) show another result: the interaction between institutional incentives and voluntary participation can take off social traps, at the same time with hiking up cooperation. The most important result of this work is the highlight of a long-run effect: social learning will lead to a cooperative society, irrespective of the number of free-riders and cooperators playing at the beginning. The recent paper by Dercole et al. (2013) describes the effect of moderate punishment. They show that it shrinks the initial conditions as well as driving towards the fixation of cooperation. The authors' conclusion is that over-punishment is not needed, and equilibria characterized by cooperation can be obtained with a gentle punishing scheme.

In their recent work Solferino and Taurino (2015) investigate the possible evolution of cooperation when you have individuals not eager to cooperate initially, but willing to "get back in the game" later on. They want to participate and cooperate for the common good in a second time. Authors show that if the other players are in turn willing to give them a second chance, then the "early stage defectors" will establish cooperation forever. On the other hand, if they meet the defectors, they will support only a cost at the second stage

and then the cooperation fails in the long run. An example of this case is the conviction as punishment for those who are redeemed to have the opportunity, after their penalty, to reenter society and cooperate for the common good.

In this work, we aim to add a contribution to this new strand of the recent literature on cooperation and punishment. We aim to investigate the probabilities of a stable cooperation in an environment where the agents take into consideration the others' behavior to achieve its goal. In particular we extensively apply the analytical results of the traditional Q-learning Model developed by Kyanercy et al.(2012) in a context of punishment and cooperation. In a Q-learning Model, people learn strategies based on the value of the related action itself and the possibly expected reward.

Xie M.C. and Tachibana, A. (2007) focus their work onto "trash pickup". They show the behavior of agents interacting with the environment and learning how to perform a task (trash collection) as well as acquiring a cooperative behavior. With this purpose, the authors develop a Q-learning model as a representative technique of reinforcement learning.

Waltman L. and Kaymak U. (2008) present a Q-learning Model to understand firms behavior in a repeated Cournot oligopoly game. Their results show how in a situation with no punishment and no explicit communication, firms tend to collude with each other.

In this work, we try to point out that when subjects have strong intrinsic motivation from achieving a certain action, then cooperation can remains rather stable or being the preferred action in the long run even if the others subjects do not cooperate. This mechanism is the case of gift and strong unconditional reciprocity.

Nevertheless when these intrinsic motivations are low, there are still rooms for cooperation by applying the reinforcement learning strategies, depending on the use of strategic measures based on punishment related to the free-riding realized.

We demonstrate how the long-term learning process, combined with appropriate sanctions in the context of strategic adoption, can open the range of network topologies. This openness will guarantee the development of cooperation in a wider range of costs and temptations. Our results suggest that a balanced duo of learning and punishment may help to preserve cooperation when there are not enough intrinsic motivation or utilities from cooperating. Cooperation is hence a "habbit" that can be taught (and learned) whether or not there are

intrinsic motivations.

Our results show that: i) if it is possible to free-ride and not being adequately punished, there will always be an incentive to deviate from the cooperation (e.g. the reduction of sentences are counterproductive); ii) conversely, even if the others did not cooperate, it is still possible to have someone who cooperates in any case. This possibility happens when individuals are pushed by strong intrinsic motivation, even if the rewards and fees are inadequate; iii) cooperation can be a learning process. It is possible to trigger a learning process that leads individuals to be equally cooperative, with probability greater than $\frac{1}{2}$. This process happens much more easily; the most responsible individuals are. It also depends on proper punishment.

2 The Model

2.1 The basic set-up

The Reinforcement Learning models demonstrate as repeated interactions with the environment will allow the learning of almost optimal behavior by agents.

Every interaction with the environment implies the agent makes a contingent choice, namely a choice based on the state of the environment at that particular time. Also, each choice corresponds to a reinforcement signal or a prize; that rewards the agent for the action taken. It follows that each agent has the objective of long-term learning of behaviors that allow the increase of cumulative rewards.

There are different types of implementation of the adaptation above mechanisms. Among these types, in this paper we consider the so-called Q-learning Model, where the agents' strategies are parameterized through Q-functions that characterize the relative utility of a particular action. As the Q-functions are renewed at every interaction, the agent has with the environment. In this way, there is the reinforcement of those actions producing higher recompenses. Specifically, assume only two existing actions to the agent, $i = 1, 2$. Here 1 is the cooperative choice (e.g. recycling, taking action on the environment, participate in a human rights campaign), and 2 is the non-cooperative choice (e.g. recycling but the plastic;

ignore a call to action on the environment; sign but not active participate in a human rights campaign). Let $Q_i(t)$ denote the Q-value of the corresponding action at time t. Then, after the selection of action 1 at time t, the corresponding Q-value is updated according to:

$$Q_1(t+1) = Q_1(t) + \alpha[r_1(t) - Q_1(t)].$$

where $r_1(t)$ is the observed reward for action 1 at time t, and α is the learning rate. On the contrary, if at time t the agent will select action 2 the corresponding Q-value will be updated according to

$$Q_2(t+1) = Q_2(t) + \alpha[r_2(t) - Q_2(t)] + \alpha(\beta - \phi)r_2(t)$$

where $r_1(t)$ is the observed reward for action 2 at time t, ϕ represents the penalty for not cooperating, β the percentage of the return onto the "common good" (e.g. wider rights for all; better environment etc.) one individual will benefit thanks to the investment made by those ones making "good choices" allowing for the "common good" to be realized.
β and ϕ are measured as a percentage α on the $r_2(t)$ return.
Moreover, we assume that $Q_1(t) \geq Q_2(t), r_1(t) \geq r_2(t)$ and also that if both individuals decide not to cooperate, the "common good" cannot be achieved and therefore its return is null. It is to be pointed out that $\alpha\beta$ can be seen as an extra benefit, coming from the reinvestment at rate α of the share deriving from free-riding. For example, if I decide to take action in a campaign on the environment but not to fund it and the result of such a campaign will be a better and cleaner seaside near my house, I will take all the benefits arising form a better environment without the costs. Here we focus on Boltzmann action selection mechanism (Kianercy et all.,2012), where the probability x_i of selecting the action i is given by

$$x_i = \frac{e^{\frac{Q_i(t)}{T}}}{\sum_{k=1}^{2} e^{\frac{Q_k(t)}{T}}}, \quad i = 1, 2 \tag{1}$$

where the *temperature* $T > 0$ controls the individual's exploration/exploitation tradeoff.
Into the next sections, the model will be used to analyze one agent's decisions on to cooperate or not, both when other agents' behavior is considered as exogenous and if the two interact with each other.

2.2 The model with one agent.

We are interested in the continuous time limit of the above learning scheme. Toward this end, we divide the time into intervals τt, replace $t+1$ with $t+\tau t$ and α with $\alpha\tau t$. Next, we assume that within each interval τt, the agent samples his actions, calculates the average reward r_i for action i, and applies (1) at the end of each interval to update the Q-values. In the continuous time limit $\tau t \to 0$, one obtains the following differential equations describing the evolution of Q values:

$$\dot{Q}_1(t) = \alpha[r_1(t) - Q_1(t)], \tag{2}$$

$$\dot{Q}_2(t) = \alpha[r_2(t) - Q_2(t)] + \alpha(\beta - \phi)r_2(t). \tag{3}$$

Next, we would like to express the dynamics in terms of strategies rather than the Q-values. Toward this end, we differentiate x_1 in (1) with respect to time and divided by x_1, and using (2) and (3) we get:

$$\frac{\dot{x}_1}{x_1} = \frac{\dot{Q}_1(t)}{T} - \frac{\sum_{k=1}^{2} e^{\frac{Q_k(t)}{T}} \cdot \frac{\dot{Q}_k(t)}{T}}{\sum_{k=1}^{2} e^{\frac{Q_k(t)}{T}}} =$$

$$\frac{\alpha[r_1(t) - Q_1(t)]}{T} - \frac{e^{\frac{Q_1(t)}{T}} \cdot \frac{\alpha[r_1(t)-Q_1(t)]}{T} + e^{\frac{Q_2(t)}{T}} \cdot (\frac{\alpha[r_2(t)-Q_2(t)]}{T} + \frac{\alpha(\beta-\phi)r_2(t)}{T})}{\sum_{k=1}^{2} e^{\frac{Q_k(t)}{T}}}.$$

Rescaling the time, $t \to \alpha t/T$, and after some steps we arrive at:

$$\frac{\dot{x}_1}{x_1} = r_1(t) - \sum_{k=1}^{2} x_k r_k(t) - x_2(\beta - \phi)r_2(t) - Tx_2\left(\frac{Q_1(t)}{T} - \frac{Q_2(t)}{T}\right).$$

Since

$$\frac{Q_1(t)}{T} - \frac{Q_2(t)}{T} = \log e^{\frac{Q_1(t)}{T}} - \log e^{\frac{Q_2(t)}{T}} = \log\left(\frac{e^{\frac{Q_1(t)}{T}}}{e^{\frac{Q_2(t)}{T}}}\right) = \log\frac{x_1}{x_2}$$

by substitution, we finally get

$$\frac{\dot{x}_1}{x_1} = \left[r_1(t) - \sum_{k=1}^{2} x_k r_k(t) - x_2(\beta - \phi)r_2(t)\right] - Tx_2 \log\frac{x_1}{x_2} \tag{4}$$

The term in bracket square in (4) shows that the probability of taking action 1 increases with a rate proportional to the overall efficiency of that strategy. This increase is as bigger

as higher is the penalty and lower is the free-riding. Instead the second term characterizes the agent's tendency to randomize over possible actions.

Proposition 2.1. *The possibility of paying no adequate penalty, in the case of free-riding, associated with any benefit, makes the temptation to deviate from the cooperative strategy impossible to remove. This is regardless of the size of the obtainable benefit and of the utility derived from the non-cooperative behavior.*

Proof. To compute the steady state we assume $\dot{Q}_1(t) = 0$ and $\dot{Q}_2(t) = 0$.
Hence it follows that $Q_1 = r_1$ and $Q_2 = (\beta - \phi + 1)r_2$.
Therefore

$$x_1^s = \frac{e^{\frac{r_1}{T}}}{e^{\frac{r_1}{T}} + e^{\frac{(\beta-\phi+1)r_2}{T}}}$$

As we have assumed, the probability of cooperating increases together with r_1 and with the penalty ϕ. On the other hand it decreases as r_2 and β are higher. Moreover $x_1^s = 1$ if all the benefit plus the free-riding is absorbed by the penalty (a very unrealistic case: there is never the certainty to have cooperation). □

3 To forgive seventy times seven:intrinsec motivation and long-run cooperation

Consider a case similar to the above, but where the agent retains the memory of the action the other agent has taken in the period immediately before. Note the last is always considered exogenous. The agent imagines that the other will behave the same way in $t+1$, thus he gives probability 0 to the attainment of the reward if in the past the other agent has not chosen the corresponding strategy and probability 1 otherwise.
In this case, our model becomes:

$$Q_1(t+1) = Q_1(t) - \alpha Q_1(t) + \alpha I(t) r_1(t) + (\alpha(-\beta + \phi + 1) r_2)(1 - I(t))$$

and

$$Q_2(t+1) = Q_2(t) - \alpha Q_2(t) + (\alpha(\beta - \phi + 1) r_2(t)) I(t)$$

where

$$I(t) = \begin{cases} 1 & \text{if the other agent has cooperated in the period before} \\ 0 & \text{otherwise.} \end{cases}$$

In keeping with the section before, eventually we can obtain two possible cases:
if $I(t) = 1$ we are in the same situation as section before, while if $I(t) = 0$, then we obtain

$$\dot{Q}_1(t) = -\alpha Q_1(t) + \alpha\left(-\beta + \phi + 1\right) r_2$$

$$\dot{Q}_2(t) = -\alpha Q_2(t)$$

$$\frac{\dot{x}_1}{x_1} = \frac{\dot{Q}_1(t)}{T} - \frac{\sum_{k=1}^{2} e^{\frac{Q_k(t)}{T}} \cdot \frac{\dot{Q}_k(t)}{T}}{\sum_{k=1}^{2} e^{\frac{Q_k(t)}{T}}} =$$

$$\frac{-\alpha Q_1(t) + \alpha\left(-\beta + \phi + 1\right) r_2}{T} - \frac{e^{\frac{Q_1(t)}{T}} \cdot \frac{-\alpha Q_1(t) + \alpha(-\beta+\phi+1) r_2}{T} - e^{\frac{Q_2(t)}{T}} \cdot \frac{\alpha Q_2(t)}{T}}{\sum_{k=1}^{2} e^{\frac{Q_k(t)}{T}}}.$$

Rescaling the time, $t \to \alpha t/T$, and after some steps we arrive at

$$\frac{\dot{x}_1}{x_1} = -Q_1(t) + (-\beta + \phi + 1) r_2(t) + x_1 Q_1(t) - x_1(-\beta + \phi + 1) r_2 + x_2 Q_2(t) =$$

and we find

$$\frac{\dot{x}_1}{x_1} = x_2(-\beta + \phi + 1) r_2(t) - T x_2 \log \frac{x_1}{x_2} \tag{5}$$

Proposition 3.1. *Unfair behavior and lack of cooperation from the other, in the past, does not exclude the possibility of cooperation. It is so if utility from cooperation is high enough (i.e. strong intrinsic motivation), even if losses associated with free-riding are not sufficiently compensated through penalty to the free-rider. On the other hand, such possibility tends to zero as fast as the higher the share from free-riding.*

Proof. Assuming $Q_1(t) > 0$ e $Q_2(t) > 0$, to have a stationary state, it is necessary to have $\dot{Q}_2(t) = 0$ namely $\alpha = 0$. If so then $Q_1(t) = k_1$ and $Q_2(t) = k_2$ with k_1 and k_2 positive constants and therefore

$$x_1^s = \frac{e^{\frac{k_1}{T}}}{e^{\frac{k_1}{T}} + e^{\frac{k_2}{T}}}$$

As consequence if $k_1 = k_2$, then $x_1^s = \frac{1}{2}$, instead if $k_1 > k_2$ the possibility to have cooperation is higher. Conversely, if $\alpha \neq 0$ we do not have a stationary state. In such a case

$$Q_2(t) = c_1 e^{-\alpha t}, \quad c_1 > 0 \tag{6}$$

and

$$Q_1(t) = e^{-\alpha t}\left[c_2 + (-\beta + \phi + 1)\int \alpha e^{\alpha t} r_2(t)\, dt\right]. \tag{7}$$

with c_1, c_2 constant. From the (6), it is possible to note that the utility associated with action 2 decreases over time. In addition (7) asserts that if $Q_1(t)$ also decreases, and this is true for the high values of the free-riding, then the probability of having cooperative strategies will rise in t as long as $Q_1(t) > Q_2(t)$, otherwise cooperation fails. □

4 Do like me! Learning cooperative strategies trough free-riding's proportional punishment

In this section, we do not consider the other agent as exogenous, but we consider a game where the players interact with each other in a forward-looking context. Here every player chooses the best strategy according to the other's choices. In this type of choice strategy, penalties and share from free-riding assume a pivotal role. In keeping with the model from previous sections, with both agents playing, thus the expected payoffs of the two players can be represented by the table below. These payoffs are equivalent to those obtainable depending on the case players 1 and 2 play cooperating (i.e. C = Cooperation in the table below) or non-cooperating (i.e. D = Defection in the table below) strategy with probability x and y respectively.

	C	D
C	$(\alpha r_1, \alpha r_1)$	$(\alpha(-\beta+\phi+1)r_2, \alpha(\beta-\phi+1)r_2)$
D	$(\alpha(\beta-\phi+1)r_2, \alpha(-\beta+\phi+1)r_2)$	$(0,0)$

Therefore, in this model with two agents the rewards received depend on their joint action. In general, let A and B be the two payoff matrices: $a_{ij}\,(b_{ij}), i,j = 1,2$ is the reward of the first (second) agent when he selects i and the second (first) agent chooses j. Let x_i and y_i denote the probability of selecting the first action by the first and second agents, respectively then the expected rewards of the agents for selecting action i are as follows

$$r_i^x = \sum_{i,j=1}^{2} a_{ij} y_j, \quad r_i^y = \sum_{i,j=1}^{2} b_{ij} x_j.$$

The learning dynamics in a two-agents scenario are then

$$\dot{x_i} = x_i[(A\mathbf{y})_i - \mathbf{x}A\mathbf{y} + T_X \sum_j x_j \log(x_j/x_i)] \tag{8}$$

$$\dot{y_i} = y_i[(\beta\mathbf{x})_i - \mathbf{y}\beta\mathbf{x} + T_Y \sum_j y_j \log(y_j/y_i)] \tag{9}$$

where $(A\mathbf{y})_i$ is the i element of the vector $A\mathbf{y}$.

In that follows for the sake of concreteness we skip the index and denote with x and y the probability of selecting the first action by the first and second agents, respectively.

Proposition 4.1. *Irrespective of the values of rewards, if the propensity for the exploration T is very broad, it is then possible to reach a symmetric cooperative equilibrium with $x = y \in ((1/2), 1)$ for enough high values of ϕ or small values of β.*

Proof. In our two actions game the learning dynamics (8) and (9) become

$$\frac{\dot{x}}{x(1-x)} = (ay+b) - \log\frac{x}{1-x} \tag{10}$$

$$\frac{\dot{y}}{y(1-y)} = (cx+d) - \log\frac{y}{1-y} \tag{11}$$

where

$$a = \alpha \frac{r_1 - 2r_2}{T_X}, \quad b = \frac{\alpha(-\beta + \phi + 1)r_2}{T_X},$$

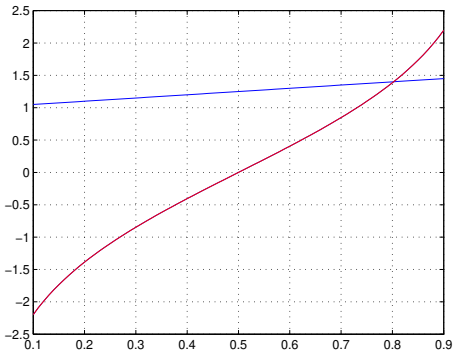
$$c = \alpha \frac{r_1 - 2r_2}{T_Y}, \quad d = \frac{\alpha(\beta - \phi + 1)r_2}{T_Y}$$

We are interested in the case of symmetric equilibria, $x = y$ and $T_X = T_Y = T$, in which case the interior rest point equation is

$$ax + b = \log \frac{x}{1 - x}. \tag{12}$$

For sufficiently large T and $b > 0$ (or you have large penalty ϕ or you have a small free-riding β), hence (12) has a unique solution $x_0 \in \left(\frac{1}{2}, 1\right)$. Graphical representation is illustrated in Fig. 1 where the blue line is the left side and the red curve is the right side of the (12)

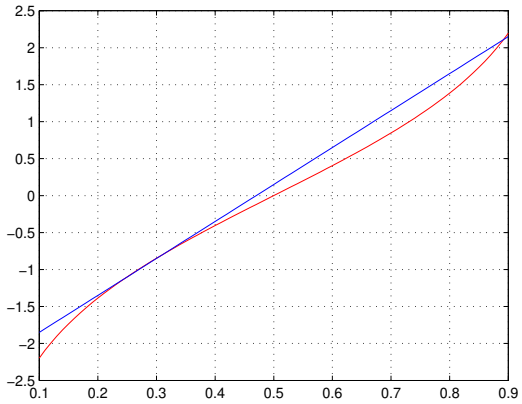
Fig.1



□

When decreasing T and $a > 0$ namely $r_1 > 2r_2$, however a second solution appears exactly at the point where the line $f(x) = ax + b$ becomes tangent to the curve $g(x) = \log \frac{x}{1-x}$.

Fig.2



Thus, in addition to (12) we should have

$$a = \frac{1}{x(1-x)} \tag{13}$$

and then it follows

$$x = \frac{1}{2}\left[1 \pm \sqrt{\frac{\alpha(r_1 - 2r_2) - 4T}{\alpha(r_1 - 2r_2)}}\right] \tag{14}$$

This solution exists only when $\alpha(r_1 - 2r_2) \geq 4T$.

Plugging (14) in (12) we find only two stable equilibrium points with

$$b_* = \log\frac{a - \sqrt{a^2 - 4a}}{a + \sqrt{a^2 - 4a}} - \frac{a - \sqrt{a^2 - 4a}}{2}, \quad b_{**} = \log\frac{a + \sqrt{a^2 - 4a}}{a - \sqrt{a^2 - 4a}} - \frac{a + \sqrt{a^2 - 4a}}{2} \tag{15}$$

We hence have two bifurcation curves (see Strogatz,2001), which meet at the cusp point $(a, b) = (4, -2)$.

Proposition 4.2. *For the rewards value $r_1 - 2r_2 > \frac{4T}{\alpha}$ it is possible to get a long run cooperative stable symmetric equilibrium with $x = y > \frac{1}{2}$, setting a penalty ϕ. Such a penalty will be higher than the share from free-riding of a quantity increasing together with T, as well as of the percentage unfairly gained from the rewards.*

Proof. From the analysis above it is possible to see how we can incentive a long-run cooperative equilibrium with probability $x = y > \frac{1}{2}$, thus staying on the stable path of the bifurcation, path defined by b_{**}.

In such a case it has to hold that

$$\frac{\alpha(-\beta+\phi+1)r_2}{T} = b_{**}$$

from that it follows

$$\phi = b_{**}\frac{T}{\alpha r_2} + \beta - 1.$$

Therefore, it seems necessary to threaten a penalty higher than the convenience arising from free-riding for an add-on as great as high is the propensity to exploration. □

5 Conclusions

Intrinsic motivation is a powerful driver of human behavior towards cooperation and reciprocity. Intrinsic motivation can not only foster cooperation, but also provides it to stay stable over the long run, even in presence of defectors. Andreoni (1989 and 1990) has described a peculiar form of intrinsic motivation as a "warm glow effect". A sort of impure altruism motivating people with a utility perceived from the sole act of giving - a positive emotional feeling they receive from the good action undertaken.

Some empirical works in a game context (Becchetti et al., 2015) show that reciprocity is positively correlated with this kind of intrinsic motivation by analyzing the level of satisfaction of participants in the context of Vote With the Wallet game. However, even if people do not have a high level of intrinsic motivation, there is possibility to boost cooperation and positive reciprocity by adopting learning strategies. With this aim, the combination, in a strategic way, of free-riding punishment and learning processes demonstrates to be effective in the long-run. Our work gives evidence to this intuitive framework, thanks to the provision of a Q-learning model in a two-players game scenario. The main point of our work is to show how the aforesaid duo of punishment and learning strategies - strategically balanced - opens the network topologies, fostering cooperation in a wider range of costs and temptations. This process will inevitably happen even in the absence, or in a poor provision of intrinsic motivation and/or immediate utility from cooperating. We may say that you can always learn (and teach) how to cooperate, providing that there is adequate punishment proportional to the free-riding. It is only when combining adequate and effective penalty with strategic learning strategies, that you can have a high probability of positive reciprocity in the long run.

Our key results demonstrate how free-riding without the "risk" of punishment represent a social possibility pushing towards uncooperative habits. This possibility may explain why the reduction

of penalties can be of no social utility.

On the other hand, the above mentioned intrinsic motivation can be the basis for unconditional cooperation. Such type of cooperative individuals will show positive reciprocity, even if rewards and fees imposed by social institutions are not adequate. Institutions can put in place social tools to develop a learning process driving individuals towards cooperation, with probability higher than $\frac{1}{2}$. Again, holding true that virtuous processes are much more easy when in presence of strong intrinsic motivation, our work shows that proper punishment is meaningful too.

References

[1] Andreoni, James, (1989),"Giving with Impure Altruism: Applications to Charities and Ricardian Equivalence," The Journal of Political Economy , 97, Issue 6, pp. 1447-14.

[2] Andreoni, James, (1990),"Impure Altruism and Donations to Public Goods: A Theory of Warm Glow Giving," Economic Journal, 100, pp.464-477.

[3] Antoci, A., Sabatini, F. and Sodini, M. (2014), "Online and Offline Social Partecipation and Social Poverty Traps: Can Social Networks save human relations?," *CRENOS WP* 2014/04.

[4] Becchetti, L., Federico, G. and Solferino, N. (2011), "What to do in globalised economies if global governance is missing? The vicarious role of competition in social responsibility, competition in social responsibility," *International Review of Economics* 58(2): 185-211.

[5] Becchetti, L., Palestini, A., Solferino. N. and Tessitore, M.E, (2014), "The Socially Responsible Choice in a Duopolistic Market: a Dynamic Model of Ethical Product Differentiation," *Economic Modelling*, 43, December, 114-123. boomze, I. (1983), "Lotka-Volterra equations and replicator dynamics: a two-dimensional classification," *biological Cybernetics*, 48: 201-11.

[6] Becchetti, L., Pelligra V. and Taurino S.F., (2015), "Other regarding preferences and betrayal aversion: insights from experimental findings and satisfaction data.", Unpublished Work.

[7] Boyd, R. and Richerson, P.J. (1988), "The evolution of reciprocity in sizeable groups," *Journal of Theoretical biology*, 132: 337– 356.

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

[9] Brandt, H., Hauert,C. and Sigmund,K. (2006), "Punishing and abstaining for public goods," *Proceedings of the National Academy of Sciences*, 103: 495-497.

[10] Bruni, L.(2006),*Reciprocità. Dinamiche di cooperazione, economia e società civile* Mondadori Eds.

[11] Dawes, R.M.(1980),"Social dilemmas," A*nnual Review of Psychology*, 31: 169-193.

[12] Dercole, F., DeCarli, M., Della Rossa, F. and Papadopoulos, A.V. (2013), "Overpunishing is not necessary to fix cooperation in voluntary public goods games," Journal of Theoretical biology, 324, pp.70-81.

[13] Egas, M. and Riedl, A. (2008), "The economics of altruistic punishment and the maintenance of cooperation," *Proceedings of the National Academy of Sciences*, 275: 871-878.

[14] Fehr, E. and Gachter, S. (2000), "Cooperation and punishment in public goods experiments," *American Economic Review*, 90: 980-994.

[15] Fehr, E. and Gachter, S. (2000), "Altruistic punishment in humans," *Nature*, 415: 137-140.

[16] Fowler, J.H. (2005), "Altruistic punishment and the origin of cooperation," *Proc.Natl. Acad. Sci.*, 102: 7047-7049.

[17] Hardin, G. (1968), "The tragedy of the commons," *Science*, 162: 1243-1248.

[18] Hauert, C. and Schuster,P. (1998), "Extending the iterated prisoner's dilemma without synchrony," *Journal of Theoretical biology*, 192: 155-166.

[19] Hauert, C.,Traulsen, A., brandt, H., Nowak, M.A. and Sigmund, K. (2007), "Via freedom to coercion: the emergence of costly punishment," *Science*, 316: 1905-1907.

[20] Hauert, C., Traulsen, A., brandt, H., Nowak, M.A. and Sigmund, K. (2008), "Public goods with punishment and abstaining in finite and infinite populations," *biol.Theory*, 3: 114-122.

[21] Kagel, J. Roth, A. (1997), *The Handbook of Experimental Economics* , Princeton, NJ: Princeton University Press.

[22] Kianercy, A., Galstyan, A.,(2012)"Dynamics of Boltzmann Q learning in two-player two-action games", *Physical Review E*, 85:041145.

[23] Nakamaru, M. and Dieckmann, U. (2009), "Runaway selection for cooperation and strict-and-severe punishment," *Journal of Theoretical biology*, 257: 1-8.

[24] Sasaki, T., brannstrom, A., Dieckmanna, U. and Sigmund, K. (2012), "The take-it-or-leave-it option allows small penalties to overcome social dilemmas," *Proceedings of the National Academy of Sciences*, 109: 1165-1169.

[25] Schuster, P. and Sigmund, K. (1983), "Replicator dynamics," *Journal of Theoretical biology*, 100: 533-538.

[26] Sigmund, K., DeSilva,H., Traulsen,A. and Hauert,C. (2010), "Social learning promotes institutions for governing the commons," *Nature*, 466: 861-863.

[27] Strogatz,S.,H. (2001)*Nonlinear Dynamics And Chaos*. West-view Press.

[28] Waltman L. and Kaymak U. (2008), "Q-learning agents in a Cournot oligopoly model,"Journal of Economi Dynamics and Control 32(10):3275-3293 ·

[29] Xie M.C. and Tachibana, A. (2007), "Cooperative Behavior Acquisition for Multi-agent Systems by Q-learning," Foundations of Computational Intelligence, 2007. FOCI 2007.