Title: **To learn Reciprocal Altruism depend on how reinforcements are offered**

ABstract

Background,

Methodology/Principal Findings,

and Conclusions/Significance

Introduction

Altruism is a kind of behaviour by which individuals choose to favour others in detriment of their own benefit. Since Darwin’s evolution theory, which is not able to explain altruism, many proposals appeared in order to account for altruist behaviours: kin selection (ref Hamilton 1), group selection (ref) and reciprocal altruism (ref Trivers 34) among others. In the reciprocal altruism theory, the loss an individual get from being altruist returns later as part of the group gain. Thus, in the long term, being altruist become the most useful strategy. In this regard, Trivers’ theory of reciprocal altruism is able to explain how natural selection favours reciprocal altruism between non-related individuals. Perhaps the most insightful example of such behaviour is the observed among vampire bats, where individuals share blood with others who have previously shared their food (ref Wilkinson 36).

Since 1971, Iterated Prisoner's Dilemma (iPD) has been proven a useful tool to study reciprocal altruism (ref Merrill Flood and Melvin Dresher in 1950, ref Trivers 34). In the iPD, two players must choose between two possible behaviours: to cooperate or to defeat. Rewards and punishments are defined in a 2x2 payoff matrix. When the game is played indefinitely, which is its iterated version, mutual cooperative behaviour is favoured. When played once, to defeat is the best strategy (ref Doebeli & Hauert 2005). However, when the game run indefinitely, evolutionary stable strategies (ESS) emerge (ref von Neumann and Morgenstern, 1944; Nash, 1949) and, under certain constraints imposed to the payoff matrix, mutual cooperation appears as the best strategy whenever reciprocity is maintained (*Pareto Optimum*). Among a huge number of reciprocal strategies, Tit-For-Tat is one of the most simple and robust (ref Axelrod and Hamilton, 1981 23) . Tit-for-Tat is based on two simple rules: to start cooperating and to do what the other player (opponent) did in the last trial.

Among many reciprocal behaviours, reciprocity and reciprocal altruism were well documented in several species. Although cooperation is needed in order to succeed in both reciprocity and reciprocal altruism, the latter add the possibility of getting reward by defeating an opponent. Some works had assessed reciprocal altruism behaviour by means of iPD paradigm in different ways, but the experiments results revealed either low levels of cooperation [wood2016cooperation] or dependant on a treatment that enhances the cooperation preference (mutualism matrix) ([stephens2002discounting], [kefi2007accumulated], [st2009long]). Direct reciprocity, which is established between two individuals, has been observed in monkeys [11][21, 22][3] and in rats [24, 25, 26, 7]. There, while food quality seemed to impact on the cooperative behaviour, a key factor to obtain reliable cooperation levels was the opponent behaviour. In this sense, individuals tended to be more cooperative with opponents that had cooperated in the past [19, 6][12]. However, when reciprocal altruism is studied, differences between species come to light. Thus, while reciprocal altruism has been proven in monkeys (ref Parker 1977), birds and rats failed to reach high levels of cooperation, even for complex combinations of rewards and punishments in the payoff matrix [35].

The reasons why some species do not learn reciprocal altruism remain obscure. A possible explanation is that animals are not able to discriminate low contrast reward contingencies. Indeed, it was shown that rats failed to discriminate the amount of reward when the number of reward units was bigger than three [4].

Here, we used the iPD to test how payoff matrix components promote or disrupt altruistic behaviour. In that sense, positive rewards (food) have been combined with negative ones (time-out) in a game where the opponent was trained to execute a Tit for Tat strategy.

**Results**

We trained twelve rats in iPD against an opponent that plays Tit for Tat strategy. Figure 1A shows a schema of the different choices a subject can do in each trial. Thus, when the subject cooperates, it receives one pellet (R) or eight seconds timeout (S) depending on whether the opponent choice was to cooperate or to defeat. On the other hand, when the subject defeats, it receives 2 pellets (T) or four seconds timeout (P), according to whether the opponent choice was to cooperate or to defeat respectively.

The criteria for cooperation was as established as the preference for pressing C lever (cooperation) over D lever (defeat) in more than 60% of the trials for five or more consecutive sessions. Eight out of the twelve animals learned to cooperate (cooperation rate 0.86±0.05, mean±s.e.m), reaching criteria in 30±4 sessions, mean±s.e.m. In figure 1B we show the mean cooperation levels for those animals during the last twenty three sessions before reaching criteria. The inset in figure 1B shows the mean cooperation level for each animal during the last five training sessions. As a consequence of the increase in cooperation levels, the average total timeout per session decreased as training progressed (0.23±0.08, mean±sem, see Fig 1C). Due to the fact that many sequences of lever pressing can give the same amount of reward and/or timeout, independently of the cooperation level, we analysed the relationship between total reward and timeout for all animals. Thus, for each animal we compared those values with the theoretical reward-timeout function when the cooperation level was set to 60%, see figure 1D. As can be seen, the higher the cooperation levels, the larger the total reward and the lower the total timeout. We then built one Markov model for the group of cooperative animals (see Figure E) and one for the group of non-cooperative animals (see Figure 1F), averaging state and transition probabilities of individuals in each group (see Materials and Methods). It can be seen that the permanency in R state was high in cooperative animals and, whenever the animal defeats (states T and P), it returns to cooperate in most of the cases (p(c|T-1)=0.76, p(c|R-1)=0.85, p(c|S-1)=0.93, p(c|P-1)=0.87). Besides, state occupancy R=0.76 was higher than T=0.1, P=0.04 and S=0.1 (p =<1e-8, ANOVA two-way test, n=8). On the contrary, in the group of non-cooperative animals, state occupancy (T=0.25, R=0.19, P=0.33, S=0.23 and p=9.736, F=0.0035, ANOVA two-way test, n=4) and transition probabilities (p(c|T)=0.43, p(c|R)=0.38, p(c|S)=0.32 and p(c|P)=0.31)) did not evidence preference for any defined strategy, see Fig S1 supplementary materials. To discard the fact that animals had had a preference for one of the levers and, in consequence, their behaviour biased independently of the training paradigm, we selected the best four cooperators and applied a reversal procedure immediately after cooperation was reached. All animals learned to cooperate after reversal (0.87±0.04,mean±sem), see Fig. 1F.

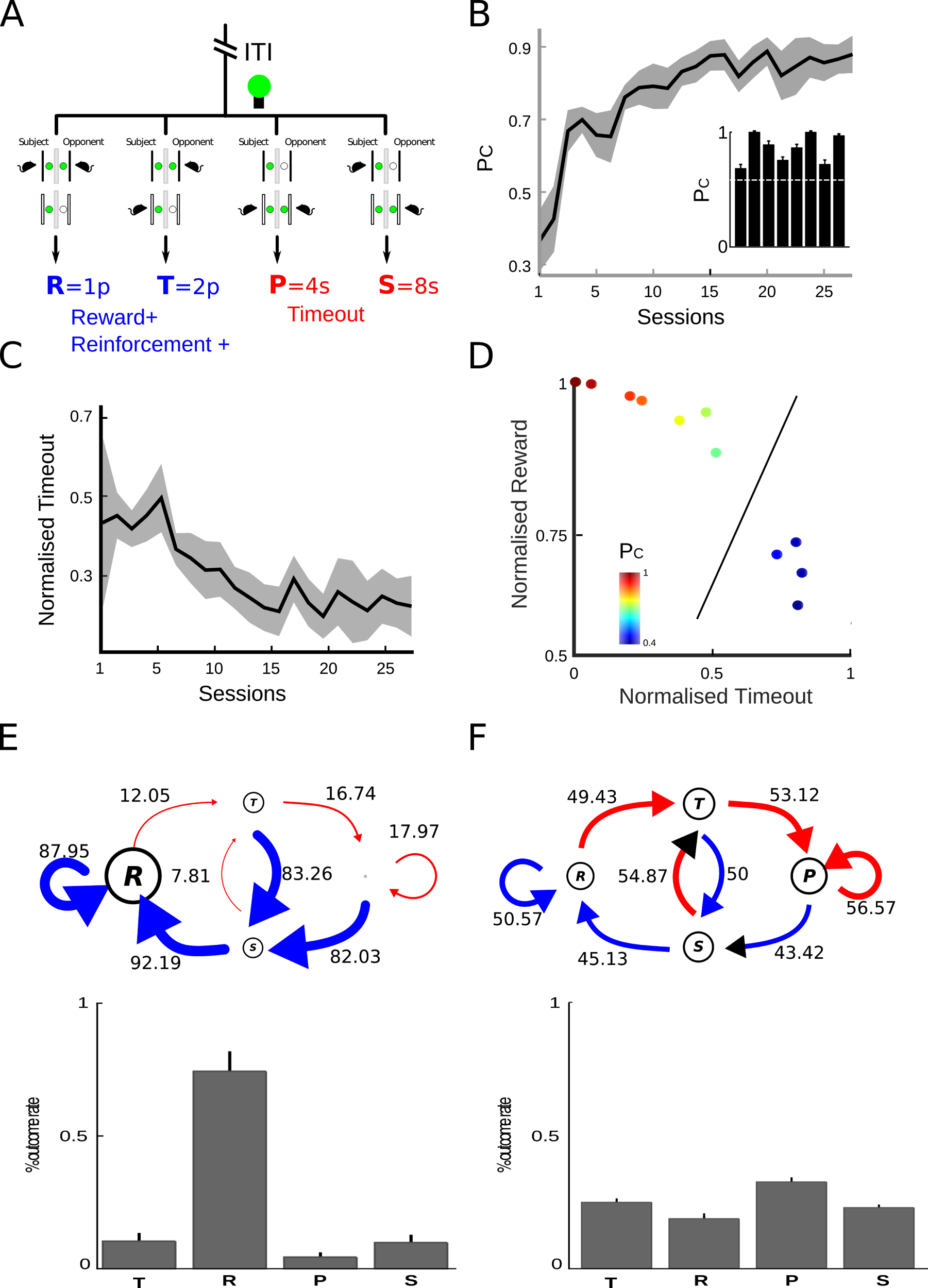


Fig.1B

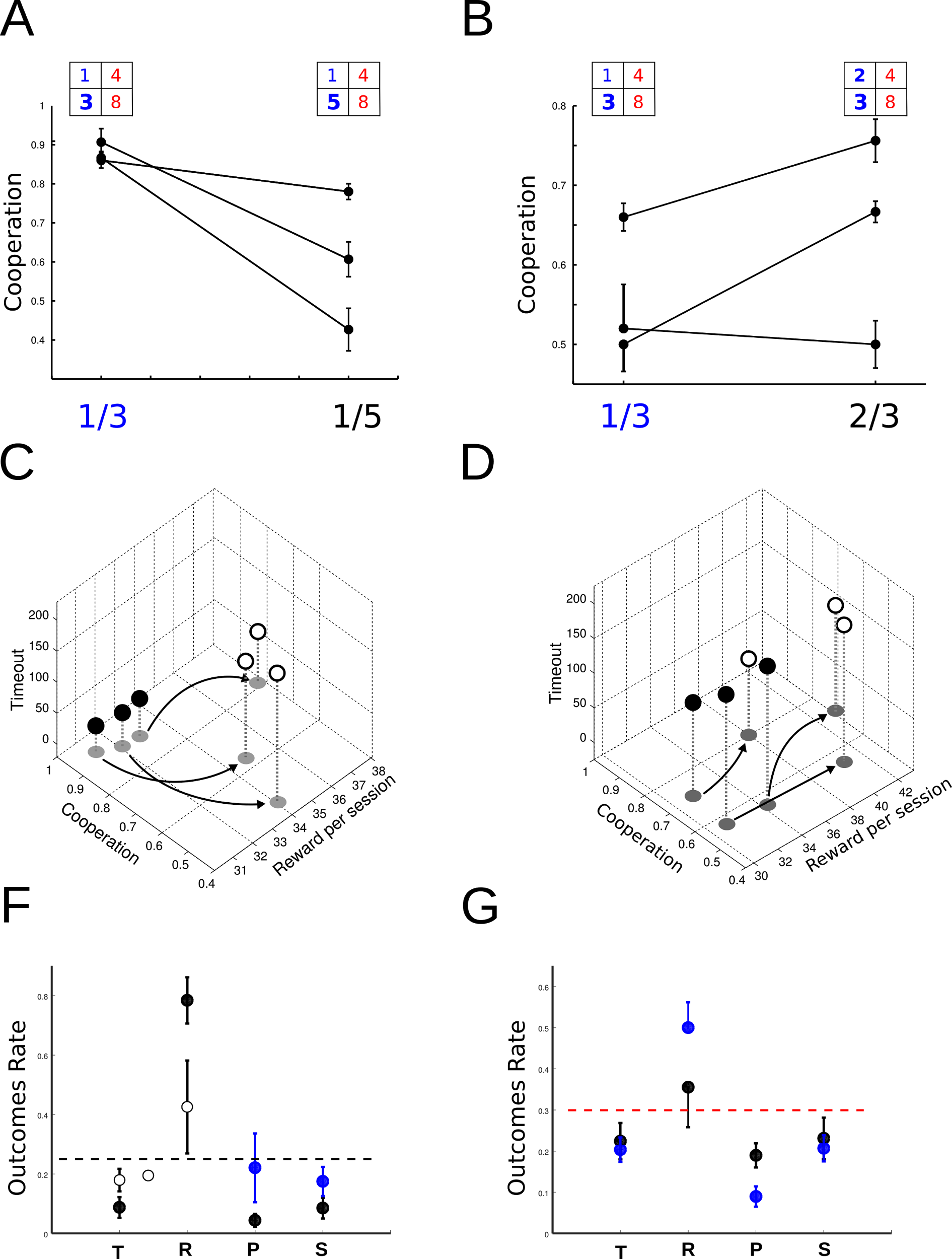
1E We performed a χ2 goodness of fit test with Bonferroni corrected over the conditional probability to cooperate and to assess the preference by the cooperated option,

We then asked how the ratio in the amount of positive reinforcement of R and T states affects cooperation learning and maintenance. We defined a contrast index CI that measures the relationship between the amount of reward in R and T as follows:

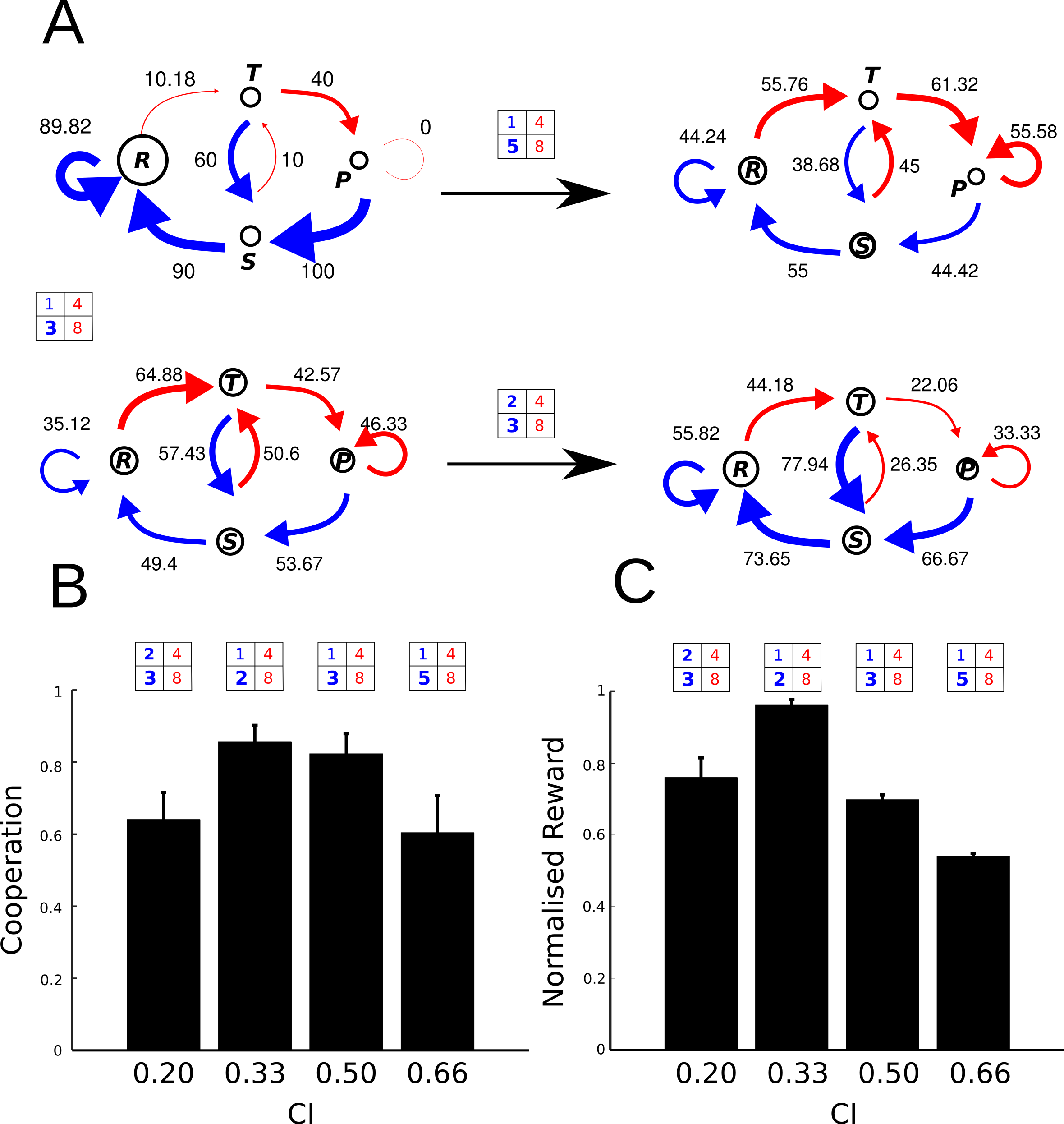
CI=(T-R)/(T+R)

Thus, in the experiment shown in Figure 1 CI was ⅓, which is the maximum contrast level constrained to a payoff matrix that favours cooperation, that is, 2\*R>T+S, assuming that S becomes a negative stimulus induced by timeout.

We trained six animals with a payoff matrix (R=1, T=3, P=4, S=8) and found that three animals learned to cooperate (0.88±0.01,mean±sem, states occupancy T=0.09, R=0.80, P=0.03 and S=0.08 ( p=<7.8-13, F=727.5, ANOVA two-way test, n=3) and p(c|T-1)=0.65, p(c|R-1)=0.90, p(c|P-1)=0.87 p(c|S-1)=0.94 (CHI2), see Figure 2A), while others did not (0.64±0.13, mean±sem, p(c|T-1)=0.47, p(c|R-1)=0.55, p(c|P-1)=0.56, p(c|S-1)=0.65 and states occupancy T=0.23, R=0.34, P=0.21 and S=0.22 and p=<0.0103, F=5.9070 , ANOVA two-way test, n=3, see Figure 2B). Then we changed the amount of reward in order to increase/decrease CI in the cooperative/non-cooperative groups. As it can be seen, a high value of CI=⅔ disrupts cooperation in Figure 2A (cooperation 0.604±0.102, mean±sem, and p(c|T-1)=0.45, p(c|R-1)=0.64, p(c|P-1)=0.62, p(c|S-1)=0.78 and states occupancy T=0.18, R=0.44, P=0.21 and S=0.17) while a lower value of CI=⅕ empowers it in two out of three animals, see Figure 2B (0.711±0.04, mean±sem, p(c|T-1)=0.62, p(c|R-1)=0.66, p(c|P-1)=0.67, p(c|S-1)=0.66 and states occupancy T=0.20, R=0.51, P=0.09 and S=0.20). We analysed how these changes in strategies impact on the amount of received reward and timeout penalties. In the group of cooperative animals, the change in T (3 pellets to 5 pellets) increased both timeout and reward, as expected when states T, P and S become more probable (pT<0.0079, pR< 0.0079 , pP< 0.0079 , pS< 0.0079, wilcoxon ranksum test), see Figure 2 C and E. It is worth noting however that the amount of received reward is not the maximum allowed, which would be delivered in the case of an animal that alternates from state T to S indefinitely. On the other hand, when we applied a matrix with a lower contrast to the group of non-cooperative animals (R=2, T=3, P=4, S=8, CI=⅕), they learned to cooperate, receiving more reward without significant changes in total timeout, see Figure 2D. In Figure 2F, we show the state occupancy probabilities for this group before and after the change in the payoff matrix. It can be seen that the occupancy probability of R state increased after the change in the payoff matrix. It can be observed a significant difference in R and P states, (pR <0.008 and pP<0.048, wilcoxon rank sum test )



From the results shown in Figures 1 and 2, it is reasonable to ask whether a fine tuning in contrasted reward encourages cooperative behaviour. We have shown that eight out of twelve (66%) animals acquired a cooperative behaviour when CI was ⅓ while three out of six (50%) succeeded when CI was ½,as expected when temptation payoff increases. In the same line of reasoning, animals that had learned cooperation under CI=½ disrupted their cooperative behaviour when CI was increased to ⅔, while those that had not learned acquired a cooperative behaviour when CI was decreased to ⅕. Figure 3A exemplifies the occupancy and transition probabilities for an animal that disrupted its cooperative behaviour when CI=½ was changed to CI=⅔. The opposite can be seen in the example of Figure 3B. A non-cooperative animal under a CI=½ became cooperative when CI was decreased to ⅕. Figures 3C and D show cooperation levels and normalised rewards. It can be seen that both variables follow an inverted U profile as a function of contrast index CI, as expected when a delicate balance between rewards at R and T is mandatory.



**Discusión**

In this work, we study the contrasted role between reinforcements in reciprocal altruist learning in rats. Traditionally, reciprocal alltruist learning is achieved by playing the iterated prisoner's dilemma (iPD game) when an experimental subject is confronted to a reciprocal opponent.

The payoff matrix used had positive and negative reinforcements with highly contrasted between pairs, positive pairs and negative pairs and also using discriminable amount of reinforcements (Capaldi Miller 1988).

In our experiment, pellets were used as positive reinforcements and timeout as negative reinforcements. This way, the positive and negative reinforcements acted as strengtheners of mutual cooperation behaviour likelihood (Skinner). To our knowledge, results show for the first time high levels of cooperation (86,11%) and mutual cooperation (76,32%) in iPD, see Fig 1B.

A dynamic system can be represented with Markov diagrams and its associated state transition vector In this case, iPD will have four possible occupancy state where experimental and opponent individual behaviour can be as follows: both cooperate (mutual cooperation, R), both do not cooperate (P), experimental subject does not cooperate when the opponent cooperates (T), and experimental cooperates when the opponent does not cooperate (S). Each state will have two associated conditional probabilities: to cooperate or not to cooperate. An individual will adopt an altruist reciprocal behaviour if when playing with an opponent with a Tit for Tat strategy, the cooperation probability is near 1, independently of the current occupancy state (T, R, P o S). If the opponent has a reciprocal behaviour, the best strategy is to return to the mutual cooperation state, R. In the first experiment in this work, we found that animals adopted two well defined strategies (fig. 1D). On one hand, a group of 8 animals proved to have learned a cooperative strategy while other 4 animals answered at random. The strategy of the first group (Fig 1E) showed cooperation probabilities according to their occupancy state T, R, P or S in 0.760, 0.845, 0.929 and 0.870 respectively, and the second group (Fig 1F) showed not significantly different from random. In various works, results were presented with Markov diagrams and its associated transition vector [viana2010cognitive, stephens2002, stevens2004, kefi2007accum] and showed that conditional probabilities of cooperation were not high when facing a reciprocal opponent.

In this protocol, with the matrix T=2p,R=1p,P=4s and S=8s, there are two theoretical strategies that maximize appetitive reinforcement: one is ALLC strategy and the other an alternating between cooperation (C) and defection (D) strategy. The latter, also maximizes possitive reinforcement when alternating between cooperation and defection, but it also increases negative reinforcement (timeout). In this case, ALLC strategy is the only one that maximizes possitive reinforcement and minimizes the negative one (Pareto Optimum). Since negative reinforcement is timeout, ALLC strategy gives more food per unit of time. In this case, the role of the negative reinforcement appears.

In order to evaluate if animals developed ALLC strategy by place preference or by reward maximization, we applied a reversion treatment, see Figure 1F, and we observed that animals re-learn reciprocal altruism when they were exposed to a new lever’s contingency.

Finally, after animals adopted a stategy, we wanted to evaluate if a change in the payoff matrix could modify their behaviour, to do so we studied the effect of modifying positive reinforcements, see Fig 2A and 2B. Animals were pre-trained with a payoff matrix where alternating between C and D strategy gives more positive reinforcements than with an ALLC strategy, and keeping a similar accumulated timeout as in the first experiment.

It was observed that only half of the animals learned to cooperate although all of them obtained the same mean amount reward (pellet), see Fig 2C, 2D.

Then, a matrix with an increased payoff T was applied to the cooperative group (Figure 2A), and we observed that cooperative behaviour decreased. Animals reduced R frequencies and increased P frequencies, proving that they prefered a small-inmediate option instead of a large-delayed option. This behaviour is similar to the one observed in birds (Stephen 1995). In the second group, we applied a matrix that keeps the proportions of reinforcements in T and R similar to the most common matrix (T=3p,R=2p). It was observed that animals modified their behaviour and became more cooperative (Fig 2B). These results show that animals that learned to cooperate with an appropriate matrix, stop cooperating when a temptation payoff (T) was enough increased (matrix with high contrast index). However, if non-cooperative animals are trained with a matrix that favours cooperation (matrix with low contrast index), they become cooperators.

In the latter case, cooperation levels achieved are comparable to results that are shared in diverse bibliography (63% of cooperation and 51% of mutual cooperation).

The main differences with other research works are the levels of cooperation and mutual cooperation achieved ([baker2002teaching, clements1995testing,

flood19832, gardner1984prisoner, green1995prisoner, marquez2015prosocial, mesterton2002economics, stephens2001adaptive, stephens2006effects, stevens2004economic, wood2016cooperation]citas, [viana2010cognitive]). A possible explanation is that animals could not discriminate among the reinforcements obtained, preventing them from learning that in the long-term the large-delayed option provides more reinforcement and consequently they did not learn iPD.

The reciprocal altruist behaviour in humans, monkeys and elefants has been studied in laboratories showing high levels of cooperation

[Wedekind1996, Kümmerli2007, de2000attitudinal, hauser2003give, Plotnik22032011], however in rats and birds those levels of cooperation were

much lower. Our results show that by using positive and negative reinforcements and an appropriate contrast index in order to favour reinforcement discrimination, rats proved to have the cognitive capacity to learn reciprocal altruism.

**Materials and Methods**

**Subject.** We used thirty male Long-Evans rats (weight 300–330g and two months old) provided by the IBYME-CONICET. Divided in two experiments, in the first one of eighteen rats, twelve experimental and six opponent, and in the second, six experimental and six opponent. Experimental subjects were housed in pairs (to allow social interaction), and opponent rats were housed individually. All rats were food restricted and maintained at 90-95% for experimental subjects, and 80-85% for opponents of free feeding body weight. and with tap water available ad libitum.The housing room was at 22ºC±2ºC and 12/12 h light/dark cycle (with lights on at 9 am). Pre-training was performed on a single standard operant chamber (MED associates Inc., USA) equipped with two stimulus light and retractable levers below the light and feeders. Also the chambers were inside an anechoic chamber with white noise (with a flat power spectral density).

The iPD experiments were performed in ad hoc dual chamber equipped with levers, lights and feeders (fig. 1A). The chambers were connected by windows in order that the rat could make olfactory and eye contact. The lever's height was 80% of maximum height of the forepaws while rearing (F. Cabrera et al., 2013). At the end of daylight, supplementary food was provided in order that rats get the amount of pellets neccesary to maintain body weight.

**Pre-experimental training.** All rats had a shaping procedure to learn the response (press a lever) to get a reinforcement (pellets). To prevent animals from choosing a lever place over the other, they learned to get reward from both sides by changing the side of conditioned stimulus. The side was changed after eight trials. All rats learned to press the correct lightng lever after four sessions.

Each rat was trained in 2 sessions per day, each trial began with the inter-trial interval (ITI) during 5 seconds, it was followed by the conditioning stimulus (light) for either 45 seconds or until a lever was pressed. One second before food is delivered, the feeder was lighted.

In the opponent's training they learned to press the lever when the light was on. In the task???Along the experiment, the side of the active lever was chosen pseudo-randomly (allowing the same side no more than four times). The opponent subject had to perform a fix ratio treatment up to FR=5 to get rewards, so as to be enough time in front of the window until the experimental subject choose a lever,

**Experiment.**

To study the reciprocal altruism in an iterated Prisoner's Dilemma game (iPD), we used a payoff matrix with positive and negative reinforcements. Positive reinforcements werepellets (Bio-Serv 45mg Dustless Precision Pellets) and negative reinforcement was timeout (a fix delay in starting a new trial). The payoff of the experimental subject was according to the matrix, and the opponent's payoff was 1 pellet when the correct lighted lever was pressed. The amount of pellets preference was previously tested on a discrimination test, showing that rats prefer 2 pellets rather than 1 pellet. We performed two sessions per day and each session had 30 trials. Each experimental subject was trained with the same opponent. The training was finished after five consecutive sessions with no changes in the cooperation rate.

The single iPD trial procedure was as follows: (1) ITI time, (2) then, the light (CS) was turned on, (3) after that both rats made their responses, the light was turned off and the reinforcement was delivered according to a payoff matrix, (4) if positive reinforcement was assigned, the feeder's light was turned on, and a second later a reward was delivered. The opponent Conditioned Stimulus (light) was controlled following a Tit for Tat strategy.The opponent received a pellet after pressing three times the lever (FR=3). If negative reinforcement (timeout) was assigned, delay time started, and the opponent subject got a pellet reward. (5) After either five seconds eating time expired or timeout was completed, a new trial started. In the first experiment the payoff matrix was: 1 pellet for mutual cooperation (R=1), 2 pellets when the experimental subject defeated and the opponent cooperated (T=2), 4 seconds of timeout for mutual defeat (P=4), and 8 seconds of timeout when the experimental subject cooperated and the opponent defeated (S=8).

At the end of the first experiment, the four rats with the best performance in cooperation were trained in a reversion treatment (see Fig. 1F).

In the second experiment we used six naive experimental rats on a different payoff matrix with greater temptation (R=1, T=3, P=4, S=8). After training, we divided rats in two groups, depending on the cooperation levels. The first group with high cooperation rate was trained with the payoff matrix (R=1, T=5, P=4, S=8) with greater temptation (T). The other group (with low cooperation rate) was trained with the matrix (R=2, T=3, P=4, S=8) that enhances cooperative behaviour (in comparison with R=1, T=3, P=4, S=8), but with low contrast between positive rewards.

All experimental procedures were approved by the ethics committee of the IByME-CONICET and were conducted according to the NIH Guide for Care and Use of Laboratory Animals.2.1 Subjects and Housing.

**Statistic.** All statistical analysis were performed using statistics library from open source software Octave and MATLAB. We pooled the data from the last five sessions where cooperation rate was stable (to calculate cooperation rate we counted the number of times a rat chose the cooperation lever per session).

We compared individual’s means of cooperation along treatment using a two-sided Wilcoxon rank sum test.

To test whether the probability of cooperation after each outcome (T,R,P or S) was different from chance (0.5), we performed a Chi-square goodness of fit test with Bonferroni corrected value of 0.05/n.

To compare mean rate of the different outcomes for each game, we performed an ANOVA two tails test. When significant α=0.05, multiple post-hoc pairwise comparative tests were performed with Bonferroni corrected value of α=0.0125.

The individual's decision rules can be described by the components of transition vectors and Markov Chain diagram. The transition vector was made up of probabilities of cooperation when the previous trials resulted in state R(reward, P(c|R-1)), T(temptation,P(c|T-1)), S(sucker, P(c|S-1)) or P(punishment, P(c|P-1)) respectively. If every component of this vector is 0.5, the agent's decision rule is random mode. Markov Chain diagram show the graphic representation of the complete decision making rule for each rat.

Supplementary Materials

Fig S1 Markov and outcome rate

Figure 1\*. Experiment 1, High level of cooperation in iPD. (A) Dual   
operant box diagram and the matrix with positive(blue) and negative(red)   
reinforcement is shown. The iPD game had four possible states: R(reward) mutual cooperation, P(punishment) mutual dessertion defection??, T(temptation) subject   
defeated and opponent cooperated and S(sucker) subject cooperated and opponent defeated. The reinforcement was used in first experiment(iPD training and reversion). The opponent´s light was used in order to perform a Tit for   
Tat strategy. (B,C) Time-course of cooperation and timeout rate along the   
last 23 games sessions. In the last 5 sessions, the mean±sem of   
cooperation was 0.86±0.05 and timeout was 0.23±0.08. (D) Total reward   
versus timeout for all animals (color bar means cooperation mean). Each   
animal was compared with the theoretical reward-timeout function when the   
cooperation level was set to 60% (black continuous line). The higher the   
cooperation levels, the larger the total reward and the lower the total   
timeout.(E). Markov Chain diagram shows the probabilities of   
transition between states (p(c|T-1)=0.76, p(c|R-1)=0.85, p(c|S-1)=0.93,   
p(c|P-1)=0.87). The arrow represents transitions: driven by cooperation in   
blue, and driven by defection in red (the arrow thickness is proportional to   
transition probability). The size of circles is proportional to the state   
occupancy ratio and below, bars show the occupancy ratio when the   
cooperation reaches stability. (R=0.76, T=0.1, P=0.04, S=0.1, p =<1e-8,   
ANOVA two-way test, n=8). Asterisks denote significant differences from   
multiple comparisons using one-way ANOVA and Bonferroni correction. (F)   
Evolution cooperation rate before and after reversion treatment. Graphs show a moving average with samples of 3 sessions (the mean and sem from   
reversion on the last five sessions was 0.87±0.04).   
  
\*Figure 2\*. Effect of changes in the amount of positive reinforcement of R and   
T. (A) The cooperation is strongly affected by change of temptation payoff   
(matrix with CI=?), decreasing when T payoff increased (significantly   
difference, p< 9.8e-06, wilcoxon ranksum test), but cooperation (B)   
enhances when the matrix favours cooperation (p<0.0062, statistics using two of three subjects, because one had no significant difference after matrix   
change, p>0.05(0.7063)). (C,D) 3D graphs related to cooperation,   
reward and timeout. In the group of cooperative animals, the change in T (3   
pellets to 5 pellets) increased both timeout and reward in order to   
decrease cooperation (p<0.05(6.7e-05)). In the group of non-cooperative   
animals, they learned to cooperate by receiving more reward without   
significant changes in total timeout (p>0.05(0.081)). (E,F) Mean±sem state   
occupancy rate graph in the last 5 sessions from cooperative and non-   
cooperative groups. Asterisks denote significant difference, after matrix   
changed, among T, R, P or S state occupancy and dash line indicates the   
level of equal rate in each state (that corresponds to a strategy with   
strongly random component).

\*Figure 3\*. (A) Over: Occupancy state and transition probabilities for an   
animal that disrupted its cooperative behaviour when CI=½ was changed to   
CI=?. Probabilities of cooperation following a previous occupancy state   
were all over the chance (before/after: p(c|T-1)=0.65/0.45,   
p(c|R-1)=0.90/0.64, p(c|P-1)=0.87/0.62, p(c|S-1)=0.94/0.78). (Below) The   
opposite situation can be seen, non-cooperative animal becomes cooperative   
when CI=½ was decreased to CI=? in a matrix that favours cooperation   
(before/after: p(c|T-1)=0.47/0.62, p(c|R-1)=0.55/0.66, p(c|P-1)=0.56/0.67,   
p(c|S-1)=0.65/0.66). (C, D) show cooperation and timeout levels as a   
function of contrast index CI. Here, it can be seen that both variables   
follow an inverted U profile in correlation with the IC contrast index   
increase and if the payoff matrix favours or not the cooperation behaviour.

Supporting