REVISTA MEXICANA DE ANÁLISIS DE LA CONDUCTA
MEXICAN JOURNAL OF BEHAVIOR ANALYSIS
28, 211-227
NÚMERO 2 (SEPTIEMBRE)
NUMBER 2 (SEPTEMBER)

# AN OPERANT APPROACH TO THE ITERATED PRISONERS' DILEMMA: EMERGENCE OF STABLE COOPERATION AFTER INDIRECT REINFORCEMENT OF CONTROLLING BEHAVIORS IN ARTIFICIAL LEARNING AGENTS

## UNA APROXIMACIÓN OPERANTE AL DILEMA ITERADO DEL PRISIONERO: EMERGENCIA DE COOPERACIÓN ESTABLE LUEGO DE REFORZAMIENTO INDIRECTO DE CONDUCTAS CONTROLADORAS EN AGENTES ARTIFICIALES DE APRENDIZAJE

**JÉRÉMIE JOZEFOWIEZ, JEAN-CLAUDE DARCHEVILLE**[1]
UNIVERSITÉ CHARLES DE GAULLE, LILLE, FRANCE

**PHILIPPE PREUX**
UNIVERSITÉ DU LITTORAL, CALAIS, FRANCE

## ABSTRACT

We propose an operant approach to the emergence of cooperation in the iterated Prisoners' Dilemma (IPD). The approach yields to the design of reinforcement-learning agents whose behavioral repertoire includes not only cooperation-related behaviors, but also controlling behaviors that may influence the behavior of the other player. The task of an agent is to learn to coordinate its own cooperation- and control-related behaviors with those of the other agents. It is suggested that this situation is closer to natural cooperative situations than the classical approaches to the IPD.

*Key words*: iterated Prisoners' Dilemma, controlling behaviors, reinforcement learning, computer agents, actor/critic architecture, Rescorla-Wagner equation, Staddon-Zhang equation

---

[1] This work was supported by a grant from the "Conseil Régional du Nord-Pas-de-Calais" (Contract n 98 49 0262). We would like to thank José E. Burgos for his invitation to contribute to this issue and for comments on an earlier draft of this paper. Corresponding author: Jérémie Jozefowiez, Unité de Recherche sur l'Evolution des Comportements et des Apprentissages, Université Charles De Gaulle, Lille, France. E-mail: jozefowiez@univ-lille.fr

learning, computer agents, actor/critic architecture, Rescorla-Wagner equation, Staddon-Zhang equation

## RESUMEN

Proponemos una aproximación operante a la emergencia de cooperación en el dilema iterado del prisionero (IPD). La aproximación lleva al diseño de agentes de aprendizaje por reforzamiento cuyos repertorios conductuales incluyen no sólo conductas cooperativas, sino también conductas controladoras que pueden influir la conducta del otro jugador. La tarea de un agente es aprender a coordinar sus propias conductas cooperativas y controladoras con aquellas emitidas por los otros agentes. Se sugiere que esta situación es más cercana a las situacones cooperativas naturales que las aproximaciones clásicas al IPD.

*Palabras clave*: dilema iterado del prisionero, conductas controladoras, aprendizaje por reforzamiento, agentes digitales, arquitectura actor/crítico, ecuación Rescorla-Wagner, ecuación Staddon-Zhang

---

The iterated Prisoners' Dilemma (IPD) is a two-player game in which, at every turn, each player must either cooperate or not cooperate with (defect from) the other. The payoffs for all the possible combinations of these two behaviors are shown in Table 1.

Table 1
Payoff matrix for the Iterated Prisoners' Dilemma (IPD)

| A2\A1 | Cooperation | Noncooperation |
|---|---|---|
| Cooperation | A1:R<br>A2:R | A1:T<br>A2:S |
| Noncooperation | A1:S<br>A2:T | A1:P<br>A2:P |

If P, R, T, and S denote four quantitatively different payoffs (e.g., amounts of money or food), then this table must satisfy two conditions in order to qualify as an IPD payoff matrix. First, $T>R>P>S$ must be the case. That is, at any turn, defection must be the best strategy (i.e., payoff must be highest) for an *individual* player if and only if the other one cooperates in that same turn. So, if a player cooperates, it should be better for the other player

to defect and 'exploit' its partner's cooperative behavior. However, mutual cooperation must return a higher individual payoff than mutual defection (i.e., $R>P$), although the latter must return a higher individual payoff than unilateral cooperation (i.e., $P>S$). Second, $2R>T+S$ must be the case. That is, the *overall* payoff for mutual cooperation must be higher than the one for unilateral cooperation (or defection).

IPD has attracted a much attention in a wide variety of disciplines, from economics to evolutionary biology. This is due to the fact that it emphasizes a surprising and important property of many cooperative situations, namely, that cooperative contingencies not only favor cooperative but also exploitative behavior. This is a threat to the long term stability of cooperation, as it has been demonstrated in laboratory with humans and animals, and in certain more natural situations (Axelrod, 1984). For instance, Green, Price, and Hamburger (1995) showed that pigeons trained in a Skinner box failed to demonstrate cooperation, even when they played against Tit-For-Tat, a computer strategy that promotes cooperation. On the other hand, real organisms embedded in a natural IPD situation are able to maintain stable cooperative interactions. Several impressive examples can be found in natural populations. One of the best known cases refers to vampire bats that share blood with unlucky conspecifics that haven't been able to collect enough blood during their nocturne hunt (Dawkins, 1989). So, how can stable cooperation emerge in an IPD?

To answer this question, researchers have designed and studied how computer strategies behave in an IPD situation, through "ecological" computer tournaments (Axelrod, 1984). In such tournaments, computer agents using different strategies interact with one another in an IPD situation. The cumulative gain of an agent at the end of a tournament in a given generation determines the number of copies of itself it leaves for the tournament in the next generation. This methodology allows for the detection of good strategies, as well as the study of complex dynamical phenomena, such as oscillations or chaos in the evolution of the population of agents (Delahaye & Mathieu, 1995). Although some of these results have been applied to the interpretation of the phylogenetic evolution of cooperation in animal populations (Axelrod & Hamilton, 1981; Dawkins, 1989), they have relied more on formal than on biological constraints.

Approaches to the IPD using reinforcement learning algorithms have only been recently proposed, which has started the exploration of a class of biologically plausible strategies that have been previously neglected. In the present paper, 'reinforcement learning' does not refer to the study of operant conditioning in experimental animal psychology, but rather the study of learning algorithms in artificial agents, based roughly on the idea of selection by

consequences. These algorithms have been inspired by experimental psychological research on animal learning (see Sutton & Barto, 1981) and constitute an active research field in contemporary artificial intelligence and artificial life. Reinforcement learning algorithms have been applied by some behavior analysts to the modeling of operant behavior (e.g. Donahoe, Burgos, & Palmer, 1993; Hutchison, 1998). An excellent introduction to reinforcement learning in artificial intelligence can be found in Sutton and Barto (1998).

Sandholm and Crites (1996) have developed a series of computer experiments using Q-learning agents, a popular reinforcement learning method that was originally developed on formal grounds by Watkins (1989). Also, Burgos (1999) has applied the Donahoe-Burgos-Palmer neural network model of operant and Pavlovian conditioning (Donahoe, Burgos, & Palmer, 1993) to the simulation of learning under an IPD situation. Reboreda and Kalcenik (1993) also devised a simple Pavlovian model based on the Rescorla-Wagner equation (Rescorla & Wagner, 1972) to interpret their data obtained with starlings playing an IPD-like game.

The present research is another example of a reinforcement-learning approach to the IPD. It differs from the previous studies not only in the architecture of the agents used in the simulation (see below) but also in the guiding hypotheses. We consider the IPD payoff matrix as a *group contingency*, that is, as a contingency of reinforcement in which the consequences of the operant do not only depend on the behavior of a given organism, but also on the behavior of other organisms embedded in the same situation (Schmitt, 1984). Experimental studies of group contingencies have demonstrated the spontaneous emergence of collateral behaviors not explicitly arranged for contingency (see Schmitt, 1984 for a review). The role of such behaviors in the success of the interaction has been suggested by Lubinski and MacCorquodale (1984), in relation to symbolic communication between pigeons. In this study, one pigeon maintained the interaction with a conspecific that was always deprived of food, no matter the physiological state of the former, due to the non-programmed emergence of species-specific behaviors in the permanently-deprived pigeon. Based on these studies, we propose that a group contingency not only reinforces behaviors explicitly specified by the contingency (let's call them 'principal behaviors'), but also collateral behaviors, insofar as they have an effect on the principal behaviors of the other organisms embedded in the group contingency. The interplay between indirectly-reinforced emergent collateral behaviors and directly-reinforced principal behaviors determines the outcome of the interaction. If we apply this analysis to the IPD, it leads to the conclusion that the payoff matrix does not only bring the direct reinforcement of cooperative and defecting behaviors (the principal behaviors), but also the indirect reinforcement of collateral behaviors. More specifically,

whatever you do as a player in an IPD situation, it is always better for you that the other player cooperates, because this will either prevent him/her from defecting or it will allow you to exploit him/her. The emergence of these controlling behaviors may explain why some groups are able to maintain a stable cooperation while others are not. Perhaps, previous studies on the IPD have overconstrained the situation by restricting the behavioral repertoire of the players. This analysis of the way a stable cooperation can emerge in an IPD situation is tested here in a computer simulation. But first, let us describe the architecture of the agents.

**Agent architecture: The RW-critic/SZ-actor model**

Our agents were built according to an actor/critic architecture (Sutton & Barto, 1998). This architecture consists of two parts, namely, a critic that computes a prediction of future reinforcement, based on the state of the environment and current reinforcement, and an actor that chooses which actions to emit, based on the state of the environment and the evaluation computed by the critic. The critic in our agents was designed after the Rescorla-Wagner (RW) equation (Rescorla & Wagner, 1972), while the actor was designed after the Staddon-Zhang (SZ) equation for credit-assignment in operant learning (Staddon & Zhang, 1991). Hence, we called our model the RW critic/SZ actor model (see Figure 1). We chose these two equations on the basis of their simplicity, relative to other models.

The SZ equation was originally proposed to explain how a given kind of response could be selected by contingent reinforcement. The equation can also account for other reinforcement-related phenomena, like the effect of reinforcement delay, superstitious behavior, and instinctive drift (see Staddon & Zhang, 1991, for further details). We consider an agent as a collection of behavioral repertoires, a behavioral repertoire being defined as a set of behaviors that are mutually incompatible at a given moment in time (or time-step or iteration). At each moment $t$, the activation level $a$ of a behavior $i$ in a repertoire is computed according to the following equation:

$$a_i(t+1) = \mu \cdot a_i(t) + \gamma(t) \cdot a_i(t) + \varepsilon \cdot (1-\mu) \qquad (1)$$

where $0 < \mu < 1$ is a short-term memory decay rate (in our simulation, $\mu = 0.4$), $\varepsilon$ is a random variable uniformly distributed over [0,1], and $\gamma$ is a reinforcement evaluation computed by the critic (see below). At each moment $t$, the behavior $i$ with the highest activation level $a_i(t)$ in a repertoire is emitted (this is the so-called 'winner-takes-all' rule). If $\gamma(t)$ is positive (see below), then (1) amplifies the differences between the activation levels of the different behaviors constituting the repertoire. Since by definition the reinforced response

is the one with the highest activation level, it will be favored over the other behaviors in the repertoire. Hence, the reinforced response will be emitted more often, even if the overall repertoire output maintains some variability due to $\varepsilon$.

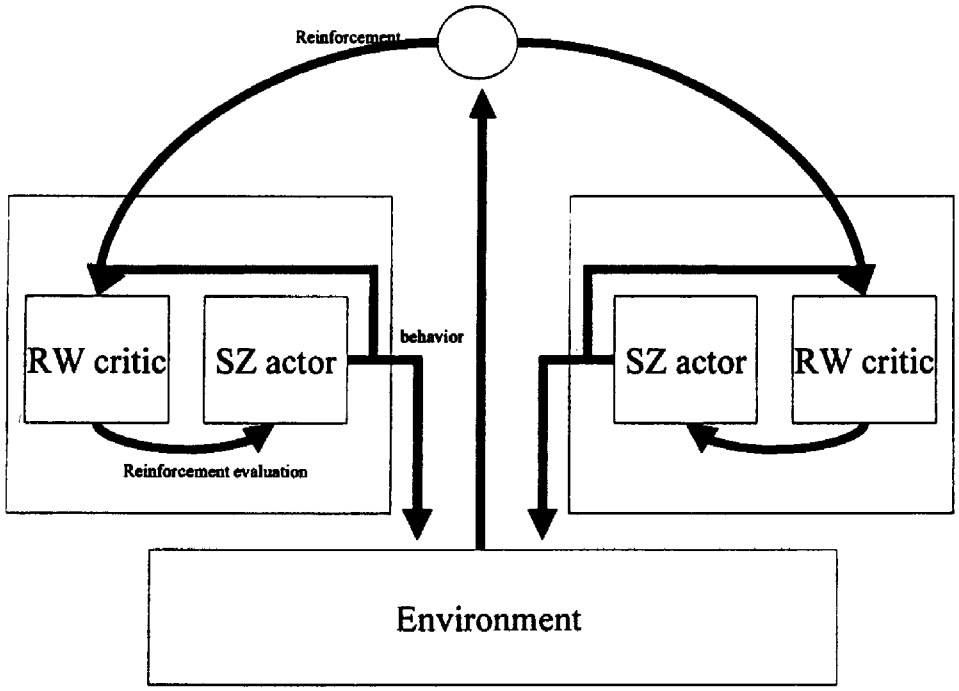


Figure 1. Schema of the RW critic/SZ actor model. An agent was a collection of RW critic/SZ actor units or agents (two in this example). A common reinforcement signal based on previous behaviors emitted by the SZ actors was delivered to the RW critics. Based on this signal and on the previous behaviors emitted by their associated SZ actor, the RW critics computed a reinforcement evaluation that was fed into the SZ actor and partially determined its output

Regarding the critic part of the model, and from the standpoint of reinforcement learning, the RW equation can be viewed as the simplest iterative prediction algorithm, for it computes a prediction of reinforcement only for the next time-step (Sutton & Barto, 1990). Each behavioral repertoire has its own RW critic that computes its own reinforcement evaluation (see Figure 2). The behavior just being emitted by its corresponding behavioral repertoire determines the state of the environment on which the RW critic bases its

evaluation. The evaluation process is summarized in the two following equations, which implement the RW equation in the context of the RW critic/SZ actor model:

$$\gamma(t)=E(t)+\rho\cdot[\lambda(t)-E(t)] \tag{2}$$

where $\gamma(t)$ is the corrected reinforcement evaluation [also used in (1)], $\lambda(t)$ is the actual amount of reinforcement given at $t$, $\rho$ is a constant learning rate (in our simulation, $\rho = 0.1$), and $E(t)$ is the overall amount of reinforcement predicted by the RW critic at $t$ (i.e., the amount of reinforcement predicted for all the repertoire behaviors occurring at $t$), which is given by

$$E(t)=\Sigma_{i=1}^{n_j} w_i(t)\cdot x_i(t) \tag{3}$$

where $w_i$ is the reinforcement expectation associated with behavior $i$, whose value depends on the agent's history, and $x_i(t) = 1$ if behavior $i$ was emitted at $t$; otherwise, $x_i(t) = 0$; and $n_j$ is the total number of behaviors constituting the behavioral repertoire $j$. The reinforcement expectation for behavior $i$ is updated according to:

$$w_i(t+1)=w_i(t)+\rho\cdot[\lambda(t)-E(t)]\cdot x_i(t) \tag{4}$$

If $x_i(t) = 1$ (i.e., if behavior $i$ was emitted at $t$) and $E(t) \neq \lambda(t)$ (i.e., if there is a discrepancy between the predicted and the actual overall amount of reinforcement), then the expected amount of reinforcement for $i$ will be changed, thus reducing the difference between E and $\lambda$ for the next time-step.

An agent can be considered as a multiagent system, each subagent being constituted by a RW critic/SZ actor (see Figure 1). Each subagent lives in its own private environment, so to speak, since the input to the RW critic is different for each subagent and they work quasi-independently, being unable to transfer information about their own functioning to other subagents. The only property they share is the actual amount of reinforcement received, $\lambda(t)$.

**Simulation of the IPD in RW-critic/SZ-actor agents**

The agents used in the present simulation had two behavioral repertoires, namely, cooperation-related (principal) behaviors and control-related (collateral) behaviors. The cooperation-related repertoire is constituted by two behaviors, namely, cooperation and defection. The control-related repertoire was also constituted by two behaviors, namely, a controlling behavior that could affect the behavior of other agents by manipulating their contingencies of reinforcement (see below, Table 2.b) and a collateral behavior that had no effect whatsoever on the behavior of other agents.

Tables 2.a and 2.b show payoffs in terms of $\lambda(t)$ values for Agent 1 (A1) as a function of A1 and Agent 2's (A2) behavior. The corresponding tables for A2 are symmetrical to Tables 2.a and 2.b, for which they are not shown here. The free parameter u represents the minimum amount of reinforcement that an agent could receive. In our simulations, u = 0.08. Table 2.a is a standard IPD matrix. The contingency described in Table 2.b slightly enhances the payoff of cooperation. In a control condition, only Table 2.a was used to compute the payoff. In the experimental condition, Table 2.b was used to compute the payoff for one agent *if and only if* the *other* agent emitted a controlling behavior. Otherwise, Table 2.a was used. Table 3 summarizes the payoffs for each agent in the control condition, while and Table 4 does the same for the experimental condition.

Tables 2.a and 2.b.
Payoffs in terms of $\lambda(t)$ values for Agent 1 (A1), according to its behavior and Agent 2's (A2). The free parameter u represents the minimum amount of reinforcement an agent could collect. In our simulations, u = .08. Table 2.a (top) is a standard, classic IPD matrix, used in the control condition. Table 2.b (bottom) was used to compute the payoffs in the experimental condition

| A2\A1 | Cooperation | Noncooperation |
|---|---|---|
| Cooperation | 6u | 10u |
| Noncooperation | u | 2u |

| A2\A1 | Cooperation | Noncooperation |
|---|---|---|
| Cooperation | 8u | 10u |
| Noncooperation | 3u | 2u |

We crossed two independent variables, namely the number of iterations in a simulation (either 1000 or 2000, which corresponded roughly to the number of iterations that the SZ equation took to settle down to equilibrium in previous simulations) and the opportunity to control. In one condition of this second variable, $\lambda$ was computed according to Table 2.a (standard IPD), so emitting a noncooperative control behavior by any of the agents did not affect differentially the contingencies, with respect to the emission of the other kind of noncooperative behavior (see Table 3). In the other condition, $\lambda$ was computed according to Table 2.b for A1 whenever A2 emitted the noncooperative controlling behavior (experimental IPD, see Table 4).

Table 3
Payoffs for each agent in the control condition (classic IPD)

| A1 | | A2 | | | |
|---|---|---|---|---|---|
| Cooperate | Control | Cooperate | Control | Payoff for A1 | Payoff for A2 |
| yes | yes | yes | yes | 6u | 6u |
| yes | yes | yes | no | 6u | 6u |
| yes | yes | no | yes | u | 10u |
| yes | yes | no | no | u | 10u |
| yes | no | yes | yes | 6u | 6u |
| yes | no | yes | no | 6u | 6u |
| yes | no | no | yes | u | 10u |
| yes | no | no | no | u | 10u |
| no | yes | yes | yes | 10u | u |
| no | yes | yes | no | 10u | u |
| no | yes | no | yes | 2u | 2u |
| no | yes | no | no | 2u | 2u |
| no | no | yes | yes | 10u | u |
| no | no | yes | no | 10u | u |
| no | no | no | yes | 2u | 2u |
| no | no | no | no | 2u | 2u |

The simulation was replicated 20 times, each time with a new dyad. For each agent in each replication, we computed a cooperation score (percentage of cooperating responses) and a control score (percentage of controlling-behavior responses). These scores were used to classify dyads into cooperative, non-cooperative, and exploitative, according to a weak criterion and a strong criterion. The weak criterion was based only on the cooperation scores. Using 50% as a cutoff point, a dyad was classified as cooperative, noncooperative, or exploitative depending, respectively, on whether its cooperation score was above this point in *both* agents, below in both agents, or above in one agent and below in the other. The strong criterion was based on both scores. A dyad was classified as cooperative according to this criterion if its cooperation and control scores were above 50% for both agents. If a dyad was classified as exploitative according to the weak criterion, and the control score for the agent with the lowest cooperation score was higher than 50%, then the dyad was classified as exploitative. The strong criterion was identical to the weak criterion for non-cooperative dyads.

Table 4
Payoffs for each agent in the experimental condition (experimental IPD)

| A1 | | A2 | | | |
|---|---|---|---|---|---|
| Cooperate | Control | Cooperate | Control | Payoff for A1 | Payoff for A2 |
| yes | yes | yes | yes | 8u | 8u |
| yes | yes | yes | no | 6u | 8u |
| yes | yes | no | yes | 3u | 10u |
| yes | yes | no | no | u | 10u |
| yes | no | yes | yes | 8u | 6u |
| yes | no | yes | no | 6u | 6u |
| yes | no | no | yes | 3u | 10u |
| yes | no | no | no | u | 10u |
| no | yes | yes | yes | 10u | 3u |
| no | yes | yes | no | 10u | 3u |
| no | yes | no | yes | 2u | 2u |
| no | yes | no | no | 2u | 2u |
| no | no | yes | yes | 10u | u |
| no | no | yes | no | 10u | u |
| no | no | no | yes | 2u | 2u |
| no | no | no | no | 2u | 2u |

The results obtained in the 20 replications are summarized in Figures 2 through 5. Figures 2 and 3 show the cooperation (left panels) and the control scores (right panels) obtained in the standard (upper panels) and the experimental IPD (lower panels), for 1000- and 2000-iteration simulations, respectively. In both simulations, cooperation scores tended to be below 50% in the standard-IPD conditions, while control scores tended to concentrate towards 50%, indicating that agents tended to emit controlling responses in a random fashion. As the upper panels of Figure 4 show, such score distributions in the standard IPD conditions resulted in most of the dyads being classified as noncooperative, a substantial number of dyads as exploitative, an none as cooperative, even by the weak criterion.

In contrast, agents tended to cooperate substantially in the experimental-IPD conditions. Indeed, as the lower left panels of Figures 2 and 3 show, a larger number of dyads satisfied the weak criterion, which resulted in a larger number of cooperative dyads in the experimental- (see empty bars in lower panels of Figure 4) than in the standard-IPD conditions.

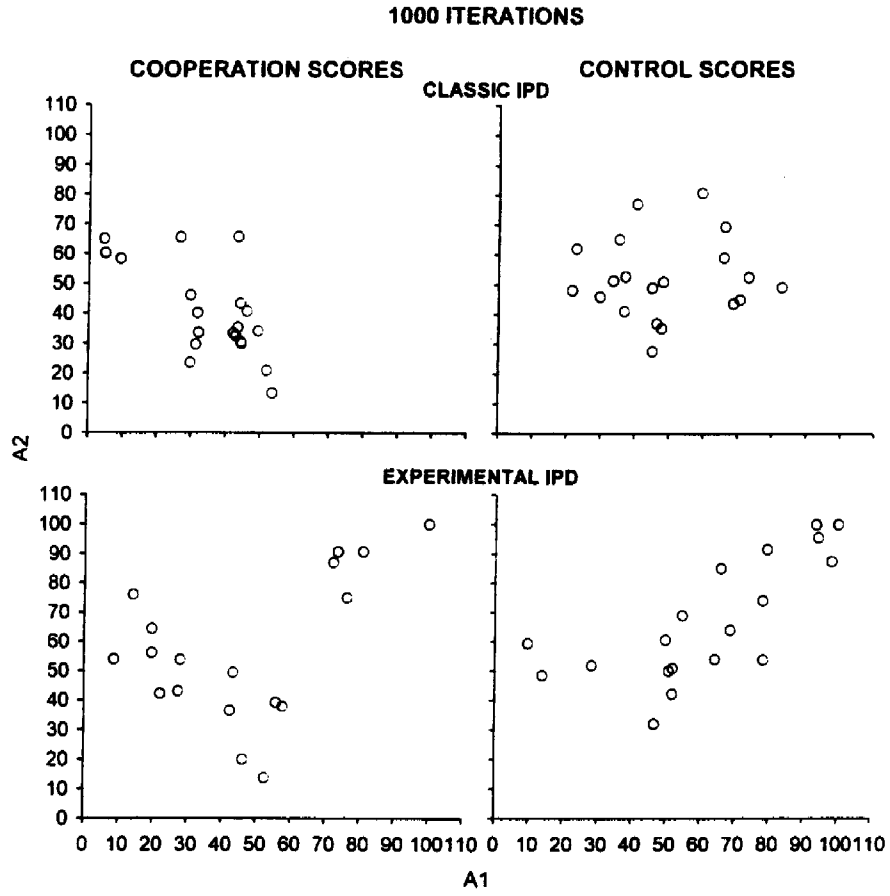


Figure 2. Cooperation (left panels) and control scores (right panels) for Agent 1 (A1) and Agent 2 (A2) in the 1000-iteration simulations with the classic- (upper panels) and experimental-IPD (lower panels) conditions

Taking the control scores into account (see lower right panels of Figures 2 and 3), we see that a larger number of dyads also satisfied the strong criterion. This resulted in a larger number of dyads that were cooperative according to this criterion (see filled bars in lower panels of Figure 4). These effects were more pronounced in the 2000- than in the 1000-iteration simulations, indicating that a prolonged exposition to the group contingencies caused agents to eventually shift from a noncooperative to a cooperative strategy. Given that exploitative dyads were substantially more numerous in

the 1000- than in the 2000-iterations conditions (an effect that is also observed, although less pronounced in the standard-IPD conditions), such a transition may have been bridged (and, to that extent, facilitated) by the exploitative strategy. Finally, an examination of the cooperation and control scores suggests that they were more closely and positively correlated in the experimental- than in the control-IPD conditions, between each other as well as between the agents within each score.

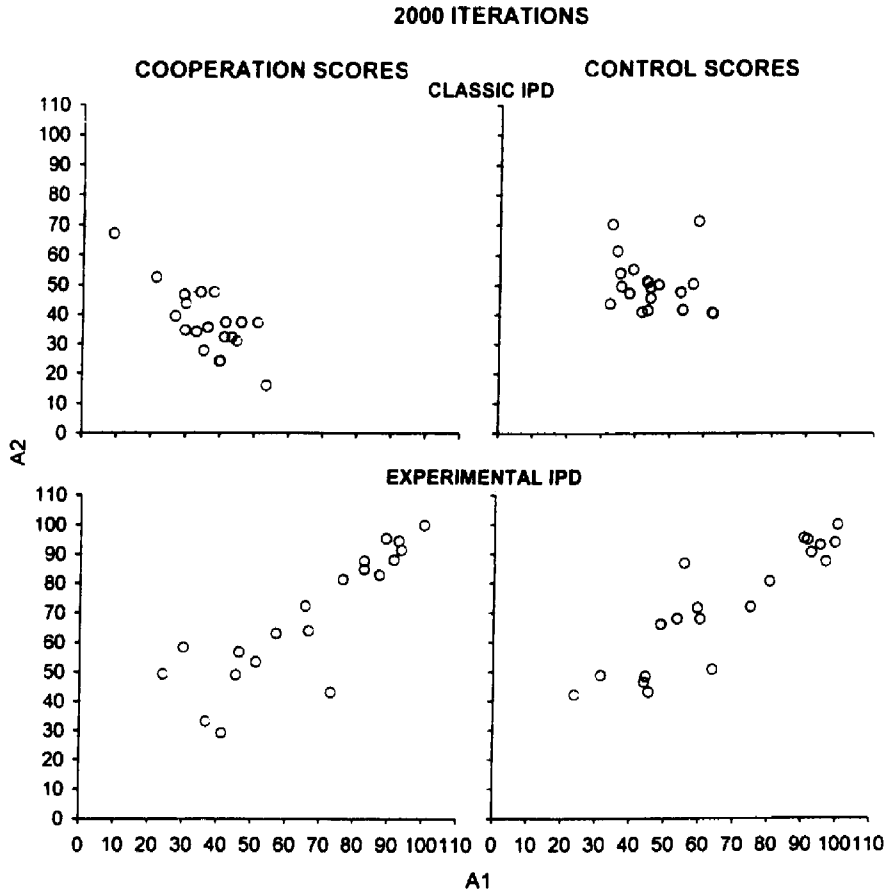


Figure 3. Cooperation (left panels) and control scores (right panels) for Agent 1 (A1) and Agent 2 (A2) in the 2000-iteration simulations with the classic- (upper panels) and experimental-IPD (lower panels) conditions

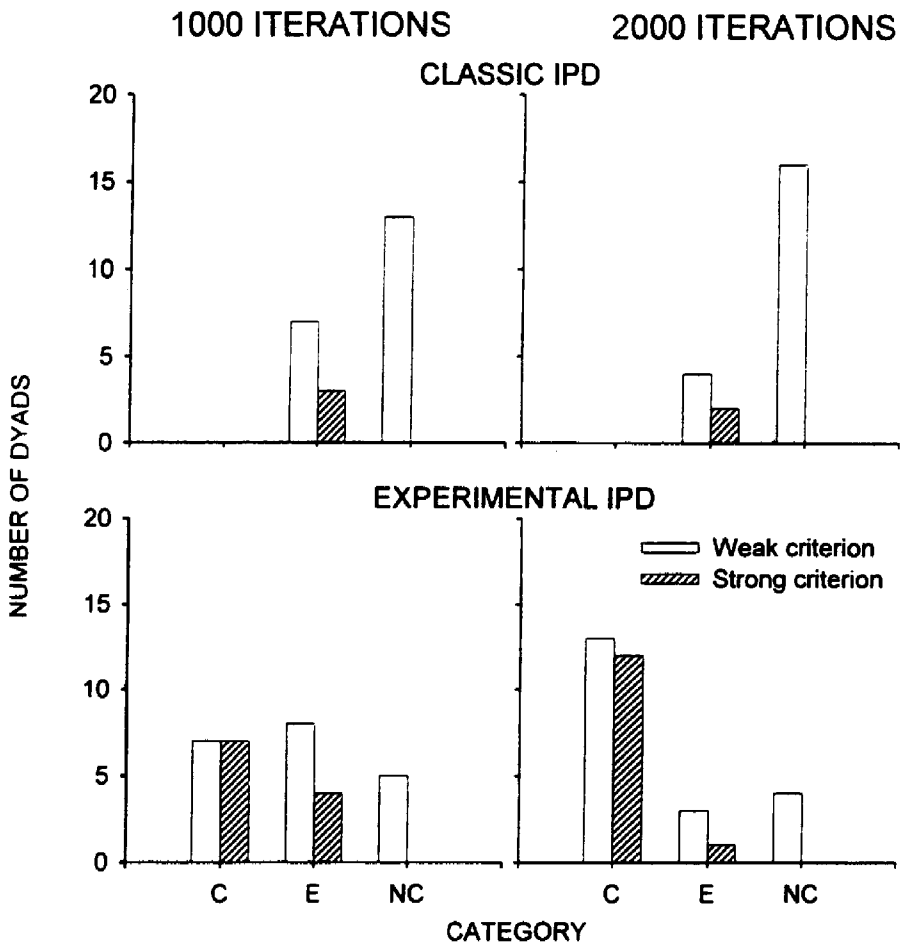


Figure 4. Classification of dyads into cooperative (C), exploitative (E), and noncooperative (NC), according to the weak criterion (empty bars) and the strong criterion (filled bars)

The above results are consistent with our analysis of the emergence of cooperation in an IPD, at least in the sense that stable cooperation emerged substantially in the experimental-IPD condition, and that such an emergence is positively correlated with an increase in control scores. However, in order to show that agents learned to cooperate *because* they learned to control each other, we must show that the increase in control scores was not superstitious. That is to say, we must ensure that such an increase was not due to an

adventitiously reinforced random biased functioning towards the control-related repertoire. To test for this possibility, we devised another simulation that was also replicated 20 times, 2000 iterations each replication. In this simulation, $\lambda$ was computed using Table 2.b only for both agents. Table 2.a was never used. Table 5 summarizes the payoff for each agent in this condition. Control scores were computed for each dyad and each agent. Results are shown in Figure 5. The distribution of control scores is far from random, but it does not display the pattern observed in the upper right panel of Figure 3. So, the increase in controlling scores observed in the experimental-IPD conditions of the previous simulation did cause the increase in cooperative scores.

Table 5
Payoffs for each agent when the contingency of Table 2.b is used unconditionally

| A1 | | A2 | | | |
|---|---|---|---|---|---|
| Cooperate | Control | Cooperate | Control | Payoff for A1 | Payoff for A2 |
| yes | yes | yes | yes | 8u | 8u |
| yes | yes | yes | no | 8u | 8u |
| yes | yes | no | yes | 3u | 10u |
| yes | yes | no | no | 3u | 10u |
| yes | no | yes | yes | 8u | 8u |
| yes | no | yes | no | 8u | 8u |
| yes | no | no | yes | 3u | 10u |
| yes | no | no | no | 3u | 10u |
| no | yes | yes | yes | 10u | 3u |
| no | yes | yes | no | 10u | 3u |
| no | yes | no | yes | 2u | 2u |
| no | yes | no | no | 2u | 2u |
| no | no | yes | yes | 10u | 3u |
| no | no | yes | no | 10u | 3u |
| no | no | no | yes | 2u | 2u |
| no | no | no | no | 2u | 2u |

Figure 5 is interesting because it suggests a way in which the agents could have handled the task. Again, a comparison between this graph and the one depicted in the upper right panel of Figure 3 reveals that control-related behaviors did not affect the amount of reinforcement earned by the agents in the standard-IPD/2000-iteration simulation, nor in the new, unconditional-reinforcement simulation. Indeed, in the former simulation, agents clearly detected the non-contingent relation between control-related behaviors and reinforcement, for which they emitted these behaviors in a random fashion. In

contrast, the same non-contingent relation in the new simulation caused agents to be superstitiously reinforced for emitting control-related behaviors. Skinner (1948) attributed the development of superstitious behavior in the pigeon to adventitious reinforcement of whatever response was occurring at the moment of reinforcement. Other behavior analysts have challenged this explanation, relying on the fact that pigeons are able to make fine discriminations between contingent and non-contingent reinforcement (e.g., Killeen, 1978). How can we explain these results?

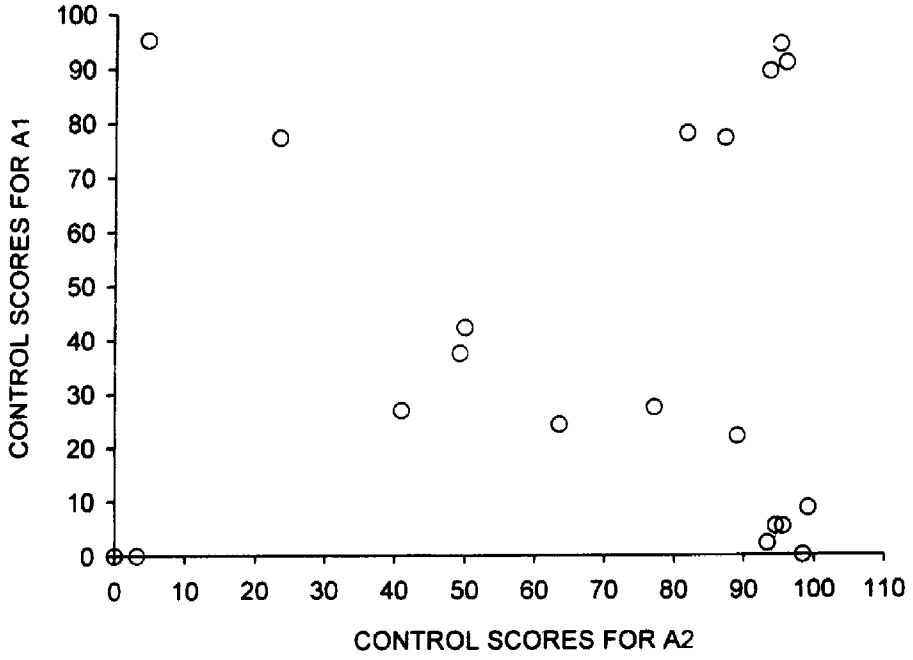


Figure 5. Control scores when the contingency of Table 2.b was used unconditionally

The RW critic/SZ actor model suggests that both explanations can be true, depending on the situation. If such a duality could be demonstrated in real organisms, this would be an interesting constraint on our models of learning. Indeed, the major characteristic of the dynamics of the SZ equation is that it converges to 0 whenever $\mu + \gamma(t) < 1$, diverging to infinity whenever $\mu + \gamma(t) > 1$. Thus, the enhancement of the differences in the activation level $a_i(t)$ is more important in the latter than in the former case, which causes the agent's behavior to become less flexible. This situation seems to be caused by

the contingency of Table 2.b but not by the one of Table 2.a. Another factor that may explain the results is the fact that the RW critic introduces a discrepancy between the real amount of reinforcement received and its effect on the SZ critic. This effect depends on the learning rate $\rho$. Rewriting Equation (4) in order to express $w_i(t)$ as a function of amplitude, reinforcement frequency, and $\rho$ may help expressing these intuitions more formally, which, in turn, will allow us to make empirical predictions about the way the RW critic/SZ actor dynamics is affected by a group contingency.

## CONCLUDING REMARKS

We have provided a behavior-analytic account of the IPD in terms of an explanation of how stable cooperation may emerge in this kind situation. A computer simulation using simple operant agents supported this analysis. Hence, we have illustrated how a group contingency can give rise to social coordination.

This work can be extended in several ways. The evolution of a whole population of agents can be studied, each agent using different controlling behaviors (some reinforcing cooperation, others defection) to see how the control-related repertoire evolves. Studies of other group contingencies, incorporating constraints from real situations that are faced by animal population can also be done. Finally, extensive comparisons with other reinforcement- learning approaches to the IPD might lead to novel perspectives in the study of this situation, and of reinforcement-learning agents in general. Extensions of the RW critic/SZ actor to other learning situations, different from the IPD one, could also be tested.

## REFERENCES

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science, 211*, 1390-1396.

Axelrod, R. (1984). *The evolution of cooperation.* New York: Basic Books.

Burgos, J. E. (1999). *Cooperation as an emergent property of selection by reinforcement in artificial neural networks*. Unpublished paper presented at the 22nd Annual Conference of the Society for the Quantitative Analyses of Behavior, Chicago, May 1999.

Dawkins, R. (1989). *The selfish gene (2nd edition)*. Oxford University Press.

Delahaye, J. P., & Mathieu, P. (1995). Complex strategies in the iterated prisoner's dilemma. In A. Albert (Ed), *Chaos and society* (pp. 283-292). Amsterdam: IOS Press.

Donahoe, J. W., Burgos, J. E., & Palmer, D. C. (1993). A selectionist approach to reinforcement. *Journal of the Experimental Analysis of Behavior, 60*, 17-40.

Green, L., Price, D. C., & Hamburger, M. E. (1995). Prisoner's dilemma and the pigeon: Control by immediate consequences. *Journal of the Experimental Analysis of Behavior, 64*, 1-17.

Hutchison, W. R. (1998). *Adaptive autonomous agent with verbal learning.* U.S. Patent #5, 802, 506.

Killeen, P. R. (1978). Superstition: a matter of bias, not detectabilty. *Science, 199*, 88-90.

Lubinsky, D., & MacCorquodale, K. (1984). "Symbolic communication" between two pigeons (*Columba livia*) without unconditioned reinforcement. *Journal of Comparative Psychology, 98*, 372-380.

Reboreda, J. C., & Kalcenik, A. (1993). The role of autoshaping in cooperative two-player games between starlings. *Journal of the Experimental Analysis of Behavior, 60*, 67-83.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variation in the effectiveness of reinforcement and nonreinforcement. In A. H. Black, & W. F. Prokasy (Eds.), *Classical conditioning II.* Englewood Cliffs, NJ: Prentice-Hall.

Sandholm, T. W., & Crites, R. H. (1996). Multiagent reinforcement learning in the iterated prisoner's dilemma. *Biosystems, 37*, 147-166.

Schmitt, D. R. (1984). Interpersonal relations: Cooperation and competition. *Journal of the Experimental Analysis of Behavior, 42*, 377-383.

Skinner, B. F. (1948). Superstition in the pigeon. *Journal of Experimental Psychology, 38*, 168-172.

Staddon, J. E. R., & Zhang, Y. (1991). On the assignment-of-credit problem in operant learning. In M. L. Commons, S. Grossberg, & J. E. R. Staddon (Eds), *Neural Network models of conditioning and action* (pp. 279-293). Hillsdale, NJ: Laurence Erlbaum.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review, 88*, 135-170.

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of Pavlovian reinforcement. In M. Gabriel, & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (pp. 497-537). Cambridge, MA: MIT Press.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction.* Cambridge, MA: MIT Press.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards.* Unpublished Doctoral Dissertation, University of Cambridge, England.