REVISTA MEXICANA DE ANÁLISIS DE LA CONDUCTA
MEXICAN JOURNAL OF BEHAVIOR ANALYSIS
26, 191-209
NÚMERO 2 (SEPTIEMBRE)
NUMBER 2 (SEPTEMBER)

SIMULATIONS OF SOCIAL BEHAVIOR: WHY AND HOW?

SIMULACIONES DE LA CONDUCTA SOCIAL: ¿POR QUÉ Y CÓMO?

SAMUEL DELEPOULLE[1]
UNIVERSITÉ CHARLES DE GAULLE,
UNIVERSITÉ DU LITTORAL CÔTE D'OPALE

PHILIPPE PREUX
UNIVERSITÉ DU LITTORAL CÔTE D'OPALE

JEAN-CLAUDE DARCHEVILLE
UNIVERSITÉ CHARLES DE GAULLE

ABSTRACT

This paper deals with an extension of behavioral principles to the study of social situations. In order to understand how individual contingencies are structured in a collective situation, we propose to investigate social situations using experiments with humans, in conjunction with simulations with behavioral artificial agents. In the first part, we present results obtained with humans in a minimal social situation. In this kind of situation, participants unknowingly interact by reinforcing and punishing each other. We observed that cooperation increased despite the fact that participants were unaware of the consequences of their behaviors, for they were not informed that they were in a social situation. The second part describes the implementation of five reinforcement-learning strategies in a computer simulation, whose performances were compared to the one observed in humans in an analogous situation. The Staddon-Zhang strategy was the best one to optimize cooperation and model human performance.

Key words: cooperation, dyads, multi-agent simulation, minimal social situation, reinforcement learning

---

[1] This research was supported by «Conseil Régional Nord-Pas de Calais» (contract n° 97 53 0283). Corresponding author: Samuel Delepoulle, Unité de Recherche sur l'Evolution des Comportements et des Apprentissages, Université Charles De Gaulle, BP 149, 59653 Villeneuve d'Ascq Cedex, France. E-mail: delepoulle@univ-lille3.fr

RESUMEN

Este trabajo es una extensión de principios conductuales al estudio de situaciones sociales. Para entender cómo las contingencias individuales están estructuradas en una situación colectiva, proponemos investigar situaciones sociales utilizando experimentos con humanos, junto con simulaciones con agentes artificiales conductuales. En la primera parte presentamos resultados obtenidos con humanos en una situación social mínima. En este tipo de situación, los participantes interactúan sin saberlo, reforzándose y castigándose unos a otros. Observamos que la cooperación se incrementó a pesar de que no estaban conscientes de las consecuencias de sus conductas, ya que no se les informó que estaban en una situación social. La segunda parte describe la implementación de cinco estrategias de aprendizaje por reforzamiento en una simulación por computadora, cuyas ejecuciones fueron comparadas con la de humanos observados en una situación análoga. La estrategia Staddon-Zhang fue la mejor en optimizar cooperación y modelar la ejecución humana.

Palabras clave: cooperación, díadas, simulación multi-agente, situación social mínima, aprendizaje por reforzamiento

---

Skinner (1953) proposed an extension of behavior analysis to social situations. This idea originated an important field of research in sociology (Homans, 1961) and experimental psychology. Many studies have investigated the effects of cooperative procedures (e.g., Hake & Olvera, 1978; Hake & Vukelich, 1972, 1973; Hake, Vukelich, & Olvera, 1975) and have compared competitive to cooperative contingencies (Olvera & Hake, 1976; Schmitt, 1976, 1984, 1986). Other social effects have been studied under a behavioral perspective, such as audit response (Hake, Vukelich, & Kaplan, 1973) or reinforcement probability (Dougherty & Cherek, 1994). In the present paper, we propose a methodology to investigate social situations, based on human experiments and simulations that make use of reinforcement-learning artificial agents.

Figure 1 depicts the method. To study a social situation, the first step (1) is to design a laboratory situation that facilitates obtaining ordered functional relations between environment and behavior. The second step (2) is to make assumptions about individual behavior that serve as explanations of the social situation of interest, specifically to determine how individual behavior is selected and what contingencies are at work in such a selection. As a third step (3), these findings are used as rules in the design of virtual agents. Finally (4), these agents are placed in a social situation analogous to the one studied with humans, in order to run a multi-agent simulation. The behavior of humans (or animals) in social situations can then be compared to the results obtained through the multi-agent simulation. If the latter matches the former, then we

have good reason to believe that the hypothesized rules of behavior constitute plausible explanations of the dynamics of the real social situation.

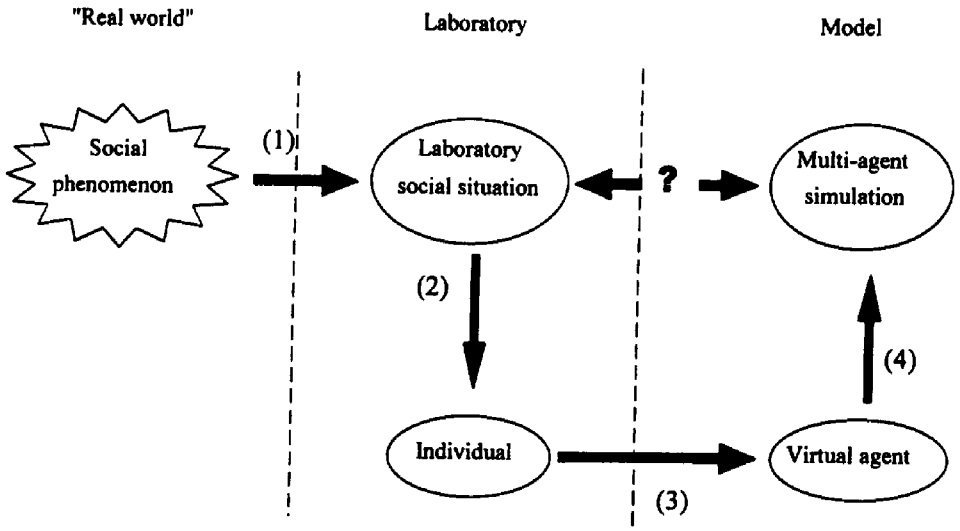

Figure 1. Diagram of the experimental/simulation process

By using this method we do not mean to say that the simulation could or should replace the experiment. Rather, our purpose is to use the simulation as a complement to the experimental analysis of behavior. Specifically, we can use the method to test different explanations of the kind of social, collective behavior observed in human dyads. Such a test would rely on comparisons between this behavior and the one observed in dyads of virtual agents. From this comparison, then, we can determine the extent to which the mechanisms that govern the simulated behavior (e.g., selection mechanisms) constitute valid explanatory hypotheses of the behavior observed in humans (or animals) in real social situations. To be sure, simulations, as simplifying, abstracting, synthesizing devices, are not supposed to mirror behavior in every single known situation. Rather, simulations, by virtue of their underlying models, are highly selective regarding particular theoretical assumptions. So, clearly, building a model of operant behavior constitutes a practical as well as a theoretical endeavor (Skinner, 1950). For our present purposes, we have chosen to design our simulated agents following a selectionist approach to operant learning (e.g., Skinner, 1938; Staddon 1983; Donahoe, Burgos, & Palmer, 1993; Donahoe & Palmer, 1994).

A social phenomenon: Cooperation

The method described above can be used to study a number of social situations. In the present paper, we concentrate on a cooperation kind of situation, which offers certain advantages for our study. Indeed, the situation represents a real social phenomenon that is observed in everyday life, as well as in the laboratory. Also, it can be defined in a relatively precise manner in terms of reinforcement contingencies. Finally, it can occur in relatively small groups (at least two individuals), which facilitates experimental-analytic work. According to Hake and Vukelich (1972), cooperation in the case of two individuals is defined by two conditions. First, reinforcement of one individual's responding must be "*at least in part dependent upon the responses of the other individual*" (p. 333). Second, the situation must allow for "*an equitable division of responses and reinforcers*" (p. 333). This second condition allows us to distinguish between cooperation and competition. Indeed, the participants in a competition situation are interdependent, just like in a cooperation situation. However, the end result in the former situation is not an equitable distribution of workload and reinforcers. In contrast, participants in a cooperation situation can improve not only their own payoffs but also the payoffs of their partners.

The earliest experimental studies of cooperation are due to Sidowski and his colleagues (Sidowski, 1957; Sidowski, Wyckoff, & Tabory, 1956). To study cooperation, they defined what they called a "minimal social situation". In this kind of situation, two individuals can mutually reward or punish one another unknowingly. Under certain conditions, Sidowski and his colleagues observed the emergence of cooperation. They suggested that their results could be explained in terms of selection by consequences. Kelley, Thibaut, Radloff, and Mundy (1962) tried to provide an experimental analysis of this selection. In order to support this latter analysis, we examined whether or not cooperation in that situation could be explained in terms of behavior selection by consequences at the level of the individual. On this basis, we used virtual agents that function according to a reinforcement-learning rule, for it is the kind of rule that has been typically used to implement behavior selection by consequences. Will cooperation emerge between such agents? To answer this question, we used the minimal social situation in order to make comparisons between the behavior dynamics of humans and that of virtual agents. To accomplish this, it was necessary to record the same kinds of variables for humans and agents alike. The next section describes the experiment with humans. The section after examines an analogous simulated situation with reinforcement-learning agents in order to compare their performance to the one observed in humans.

Human experiment: Cooperation as an exchange of reinforcement

Participants

Twenty-six undergraduate students (eight males and eighteen females) volunteered to join an experiment organized by the Psychology Department at the University of Lille III, France. They were randomly arranged into thirteen dyads and were told only that they were going to participate in a learning experiment by playing a game in a computer. They were invited individually to the laboratory and were not told that they had been paired with another individual to "play" in the experiment.

Apparatus

Two computers were placed in two separate, isolated rooms and linked through a TCP/IP network (see Figure 2). Participants were not told about the link, so they interacted with a computer without knowing that it was connected to another computer that was being used by another participant. Each computer was equipped with an Intel Pentium™ processor (one with a P120 and the other with a P90; execution of the software was not significantly different from one machine to the other), a color screen (800x600), and Windows 95™.

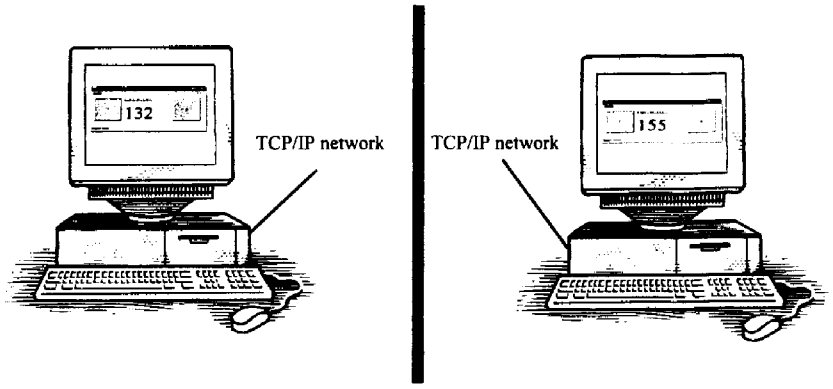

Figure 2. Experimental setup

A program was especially designed and developed in order to display stimuli on the computer screens and to record the responses emitted by each participant. The program consisted in a client/server network using Java™. The role of the clients was to display stimuli and record the participants'

responses. The server recorded the data and communicated with the clients. Each screen displayed a window of 466 by 190 pixels, containing two buttons of 100 by 100 pixels each and a counter with a height of 72 pixels and the Times New Roman font (see Figure 3). A participant could increase or decrease (without knowing it) the other participant's counter by clicking the left or the right button, respectively. A click defined a response, an increase in the counter defined a reinforcer, and a decrease defined a (negative) punishment.

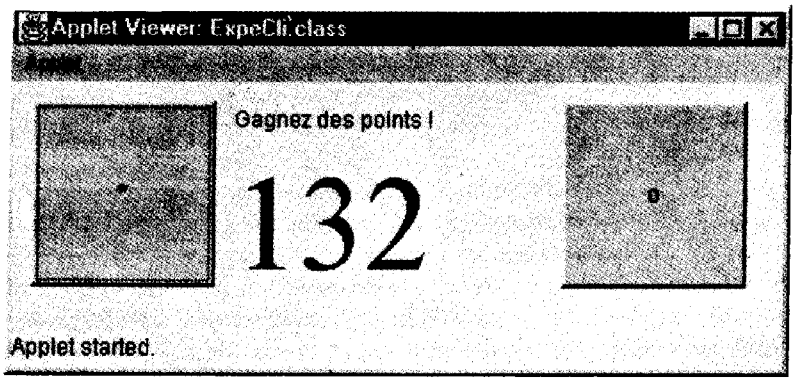

Figure 3. The client interface

Procedure

Each individual in each dyad was separately placed in a room with a computer. After asking them to sit in front it, the experimenter made sure that they were able to use the pointing device (a mouse). Then, each individual was given the following instructions: "This is a little game. You only have to earn a maximum amount of points. For this, you can use the mouse and click on the two buttons on the screen. You can click as much as you want during thirty minutes". A startup signal was given to both individuals simultaneously. Then, each individual in each dyad interacted with the apparatus (and, without knowing it, with one another through the TCP/IP link) during 30 minutes, after which they were told that the game was over. After playing the game, participants were asked to express their impressions about the game, in order to determine whether or not they had realized they had interacted with another individual. No participant reported having realized he or she had interacted with another individual during the game. In fact, participants expressed surprise after they were informed about the real purpose of the experiment.

RESULTS

For all participants, the number of responses ranged from 56 to 6638, with a median of 1446. For each dyad, a cooperation coefficient $Cc = [2R / (P+R)] - 1$ was computed, where R denoted the number of times a participant's counter increased (reinforcements) and P the number of times it decreased (punishments), due to the other participant's responding. Coefficient Cc ranged from -1, representing minimal cooperation, to +1, representing maximal competition, zero representing random responding.

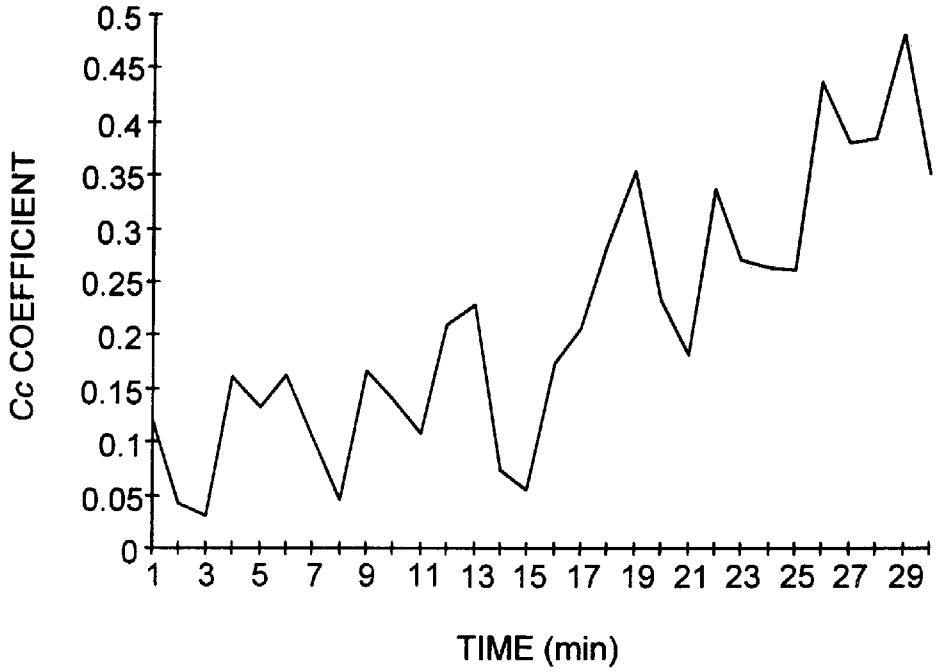

Figure 4. Changes in the cooperation coefficient during the experiment

Figure 4 shows minute-by-minute changes in Cc throughout the game, averaged across dyads. During the first half of the game, Cc was not significantly different from zero, meaning that participants tended to distribute their responses more or less evenly between the reinforcement and the punishment buttons. However, at minute 19, Cc abruptly reached .35, which was statistically significant ($t = 3.06, p < .001$). So, from that moment in the

game on, the number of reinforcements that participants gave each other increased substantially, meaning that they suddenly started to click on the left (reinforcement) button and maintained this behavior for the rest of the duration of the game. The net outcome of this behavior was a significant increase in the number of reinforcements obtained by *both* participants, indicating that they ended up (unknowingly) cooperating with one another.

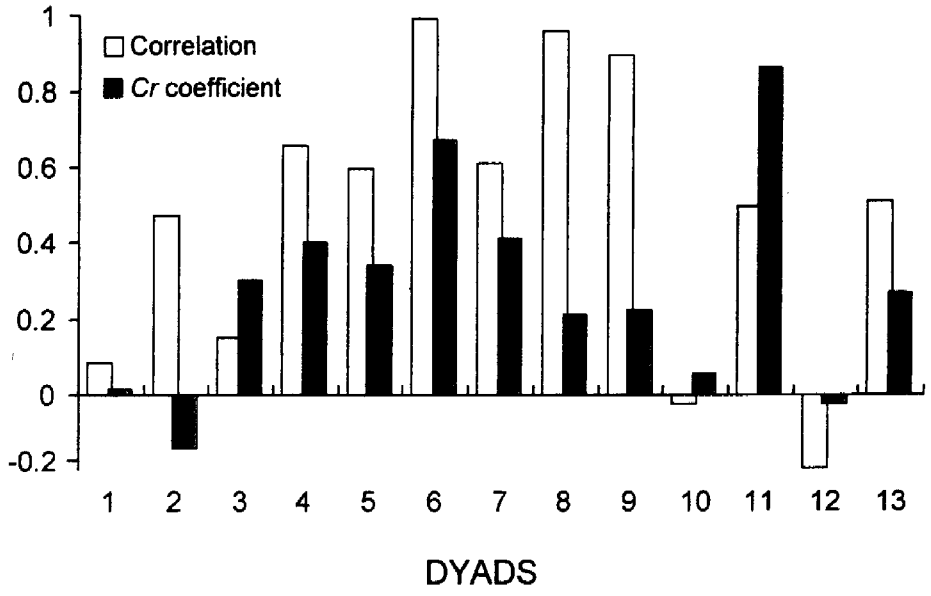

Figure 5. Comparison between the cooperation coefficient and the correlation in groups in which cooperation appears

Figure 5 shows the *Cc* coefficient (filled bars) and the correlation between the frequencies of left-button and right-button responses (empty bars), for each dyad during the last 10 minutes of the game (minutes 21 to 30). Participants in Dyads 3, 4, 5, 6, 7, 11, and 13 showed values of *Cc* near to or larger than the statistically significant value of .35, indicating a substantial amount of cooperation. The *Cc* values for Dyads 1, 8, 9, and 10 also indicate that participants exchanged more reinforcements than punishments during this period of the game, although these values were not statistically significant. Punishment was the most frequent response only in Dyads 2 and 12. Participants in these two dyads emitted a relatively low number of responses

(nearly half the number of responses emitted by participants in the other dyads).

Finally, correlation coefficients ranged from -.02 to .99, reaching statistically significant values in Dyads 4, 6, 7, 8, 9, 11, and 13, of which 4, 6, 11, and 13 also showed statistically significant *Cc* coefficients. In general, such high correlation coefficients indicate that participants in these dyads tended to respond in a relatively synchronized manner during this period of the game. The contrast between the high correlation coefficients and the low *Cc* coefficients observed in Dyads 8 and 9 indicates that participants in these dyads tended to respond in a very synchronized manner, although their levels of cooperation were not statistically significant.

DISCUSSION AND HYPOTHESIS

In a social situation where participants in human dyads responded interactively without knowing the real consequences of their behavior, cooperation increased throughout the interaction. This result is consistent with the ones reported in the field (e.g., Kelley, Thibaut, Radloff, & Mundy, 1962; Sidowski, 1957; Sidowski, Wyckoff, & Tabory, 1956). The increase observed in synchronized responding among the participants (as represented by the correlation coefficients), in those dyads where cooperation was significantly high, suggests that a cooperative contingency may have been at work in these dyads, as hypothesized by Hake and Olvera (1978). That is, once cooperative responding emerged in those dyads in which *Cc* was high, an increase in the cooperation rate of one participant reinforced the cooperative behavior of the other participant. Such a contingency would cause participants to respond in a synchronized manner, thus yielding high *Cc* as well as high correlation coefficients.

On this basis, we hypothesize that the emergence of cooperation in minimal social situations is due to behavior selection at the individual level. That is, if two participants A and B start to cooperate, then both of them are positively reinforced, which selectively increases the probability of future occurrences of the same kind of response. If A and B start punishing each other negatively (i.e., start retiring reinforcers from one another), then an extinction contingency enters in operation. This contingency, in turn, causes an initial increase in response variability and, hence, in the probability of occurrence of chance cooperative responses. Once these responses start to occur, they are selected and mutually maintained by reinforcement. Of course, it is possible that A reinforces B while B is punishing A, in which case B is likely to continue punishing A (for this behavior is reinforced by A's cooperating

behavior). However, this would constitute an extinction contingency for A, which would cause an initial, momentary increase in A's behavioral variation, thus increasing the probability of occurrence of punishing responses. The occurrence of such responses, in turn, introduces an extinction contingency for B, which, once again, would cause an increase in behavioral variation and, hence, in the probability of occurrence of cooperative responses on B's part. In this manner, synchronized cooperation, as an emergent mutual exchange of reinforcers, becomes an attractor for all possible initial states (Delepoulle, Preux, & Darcheville, 1999).

The above explanation suggests that cooperation in a dyad may develop as a result of the individual participants trying to maximize their own amount of reinforcement. In the next section, we describe simulations using reinforcement-learning agents, with the purpose of determining whether or not cooperation in these agents emerges in a similar manner. The design of the agents in question followed a particular approach to operant behavior (i.e., to behavior that is modifiable by its consequences), namely, the selectionist approach.

Virtual agents

All the agents used in the simulations described in the present section were designed after the same set of general principles. First, agents were capable of perceiving certain features of their environment. Second, they were capable of emitting behaviors that could have certain effects on their environments. More specifically, they were capable of producing or retiring reinforcers (Kaelbling, Littman, & Moore, 1996). Aside from these very general principles, however, details on how reinforcement modifies behavior, the so-called "policy", yield different models. For our present purposes, we compared, via computer simulations, the performances of five reinforcement-learning architectures, namely, the law of effect, the Hilgard-Bower rule, the Staddon-Zhang model, the action-value rule, and Q-learning. We describe their key features in the next subsections, after which we describe the simulations in question. For all the models, time was discrete and only one behavior, among a set of possible behaviors, could be emitted at any moment in time.

The Law of Effect

Thorndike (1898, 1911), in his Law of Effect, proposed that whenever an organism's response is followed by a reinforcer (i.e., by a biologically favorable consequence), the probability of future occurrences of similar responses increased. Reciprocally, when a response is followed by a negative

consequence (i.e., the absence of a biologically relevant stimulus), such a probability decreased. We propose to formalize this law in terms of the following difference equations:

$$p_i^{t+1} = \frac{p_i^t + s^t}{1+s^t} \tag{1}$$

$$p_j^{t+1} = \frac{p_j^t}{1+s^t}, \text{for all } i \neq j \tag{2}$$

where i <denotes the last emitted behavior, t denotes a moment in time, p denotes the probability of occurrence of a given response type at a given moment in time, and

$$s^t = \begin{cases} \alpha \cdot \dfrac{C_i^t}{|C_i^t|} \text{ if } C_i^t \neq 0 \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

where a denotes a learning-rate parameter and C_i^t denotes the consequence of behavior i, which can be either +1 or -1.

The sum of probabilities for all the different kinds of behaviors must be equal to one at each time t. Equation 2 is necessary to preserve this relation. This algorithm is nonlinear in that the variation rate depends on p_i. Behaviors with a low probability will change quicker than those with a high probability. In our simulations, $a = 1.5$.

The Hilgard-Bower Law

Hilgard and Bower (1975) used a rule that is very similar to the law of effect, named "linear reward-inaction algorithm". Basically, all non-reinforced actions are weakened, that is, their probabilities are decreased. This algorithm always converges with a probability of 1 on a particular action (though not always to the optimal action). The mathematical expression of this algorithm is:

if $C_i^t > 0$, then $p_i^{t+1} = p_i^t + a(1-p_i^t)$,
and for all $j \neq i$, $p_j^{t+1} = p_j^t - ap_j^t$

If action i succeeds (i.e., if $C_i > 0$), the probability is increased; if it does not, the probability remains unchanged. Like in the Law of Effect, a represents a learning rate, which was set to 0.05 in our simulations.

The Staddon-Zhang Model

Staddon and Zhang (1991) proposed a simple parallel model that aims at solving the *assignment-of-credit* problem without a teacher (unsupervised learning). They showed that this model could account for certain qualitative properties of response selection. Their model is consistent with paradigmatically normal behavioral phenomena, as well as with certain anomalous phenomena, such as autoshaping, superstition, and instinctive drift.

Each behavior is characterized by a value V_i^t, which is computed at each time-step. All values V_i compete against one another, the competition rule being "winner takes all". That is, the only behavior emitted is the one corresponding to the largest V_i^t. Two equations describe the changes in the value of V_i^t according to C_i^t:

if $C_i^t \neq 0$ then $V_i^{t+1} = \alpha \cdot V_i^t + \varepsilon^t \cdot (1 - \alpha) + \beta \cdot V_i^t$

if $C_i^t = 0$ then $V_i^{t+1} = \alpha \cdot V_i^t + \varepsilon^t \cdot (1 - \alpha)$

for all $j \neq i$, $V_j^{t+1} = V_j^t$

where $0 < \alpha < 1$ is a kind of short-term memory parameter, set to .5 in our simulations. The reinforcement parameter β should be positive for rewards and negative for punishments. In our simulations, $\beta = 1$ and ε^t was a random number between 0 and 1.

Action Value

The goal of this method is to estimate the mean payoff for each behavior and to choose the best action to follow, in order to optimize the forthcoming payoff. Sutton (1998) gives a method to compute this estimation by iteration. Let V_i^t the estimated (mean) payoff, C_i^t the immediate consequence of behavior i, and N_i^t the number of occurrences of i in the past. If behavior i is emitted, then,

$V_i^{t+1} = 1 / N_i^t \, [C_i + (N_i^t - 1) \cdot V_i^t]$,

$N_i^{t+1} = N_i^t + 1$,

for all $i \neq j$, $V_j^{t+1} = V_j^t$ and $N_j^{t+1} = N_j^t$.

The agent chooses the behavior that has produced the largest mean payoff. Such a method converges rapidly if we allow it to explore other behaviors. A simple policy could be (with N_b, the number of possible behavior)

Let e a random number in [0;1],

if e $>$ ε then i = argmax_i (V(i))

else i = random number in [1 ; N_b]

So, under this policy and with $\varepsilon = 0.1$, an agent would behave randomly.

Q-Learning

The Q-Learning algorithm was proposed by Watkins and Dayan (1992), based on the Time Derivative (TD) model (Sutton, 1988; Sutton and Barto, 1990). It is an algorithm for solving rapidly and easily stochastic optimal control problems where the agent is the controller and the environment is the system to be controlled. In order to optimize performance, Watkins introduced the quality value $Q_{s,a}$ (hence, Q-value), which represents the expected future payoff of emitting action a in state s. Q-Learning works by modifying $Q_{s,a}$ for each state-action pair, according to the following algorithm:

1. Choose an action a according the current state s. The agent enters state s' and receives the payoff r.
2. Modify $Q_{s,a}$ according to the following equation:
   $Q_{s,a} = Q_{s,a} + \alpha[r + \gamma \max_b Q_{s',b} - Q_{s,a}]$
3. Go to 1.

where α denotes a learning rate and γ denotes a discount factor. In a Markovian and stationary environment, Watkins has shown that this algorithm converges, with a probability of 1, to the optimal value. In practice, Q-Learning does not explore sufficiently the state space, for which it has been suggested to introduce variability by adding some noise to the choice performed. Instead of always emitting the best action in state s, the algorithm chooses the action at random with probability ε.

In contrast to the other four methods, Q-Learning is not "context-free". This means that its behavior is sensitive to the context. So, if a certain behavior is reinforced in a precise situation whereas the same behavior is punished in other cases, the algorithm is able to emit this behavior only in that particular situation.

$Q_{s,a}$ is a measure of the expected reward of performing the action a in the state s. This explains why the Q-Learning algorithm is able not only to emit the best behavior in a state but also to reach this state. Furthermore, Q-Learning is able to avoid states that ordinarily lead to negative consequences.

In our simulation, α and γ were set to 0.5 and $\varepsilon = 0.1$. States were determined by the variation of the counter from t-1 to t.

Simulations

Two kinds of simulations were run, namely, individual-agent and multi-agents. Each kind of simulation consisted of a total of five simulations, one for each learning algorithm described above. In the individual-agent simulations, agents were tested individually. In the multi-agent simulations, agents were tested in a situation analogous to Sidowski's minimal-social situation used in

the human experiment. Simulations were run using a Java applet. The experimenter could choose, via a graphical interface, the number of agents, the learning algorithm, and the duration of the simulation (measured in number of time-steps). The behavior of agents was displayed in a graphical window and stored in files. The simulation program can be executed from *http://www-lil.univ-littoral.fr/~delep/Expe/Simul/Agent.html*

Each individual-agent simulation lasted for 100 time-steps. At each time-step, agents could emit one out of three possible responses, namely, self-reinforcement (R), self-punishment (P), and no effect (N). These simulations were run to determine whether or not the agents could learn in a relatively simple environment. For the five learning algorithms, the evolution of behaviors was very similar, for agents succeeded in optimizing their behavior. At the end of each simulation, all agents emitted mostly R responses. Only the Staddon-Zhang agents continued emitting a significant number of N responses.

In the multi-agents simulations, an agent could emit three behaviors, namely, reinforce its partner (R), punish its partner (P), or do nothing (N). We introduced the N behavior in order to approximate better the human situation, for in this situation participants not only could respond at any moment, but also were allowed not to emit any response for as long as they wished. Applying this strategy to the simulations was critical in order to prevent the emergence of synchronized cooperation in the agents as an artifact due to the discrete and sequential way of functioning of Turing-machine based computers. We wanted such a cooperation to be due as much as possible to the close temporal-contiguity relation between consequences and responses, as stipulated by the learning algorithms. This kind of control is rarely mentioned in simulation research, despite the fact that it is critical for assessing the validity of simulation results.

In a multi-agents simulation, a given learning algorithm was applied for 1000 time-steps to 13 dyads of agents of the type stipulated by the algorithm in question. So dyads within a multi-agent simulation consisted of agents of the same type. In order to compare the performance of the agents to the one observed the participants of our human experiment, we computed a cooperation ratio $Cr = R/(P+R)$, for both humans and agents. The graphs in Figure 6 show changes in the mean dyad Cr across time, for participants in the human experiment (dotted line) and for each type of agent in the simulations (continuous line). Agent responses were grouped into 30 periods of time, each of which contained the same number of responses. During each period, each agent emitted approximately the same number of responses that a human subject did during one minute. A comparison among the graphs reveals that overall the closest simulation of human performance was achieved by the Staddon-Zhang agents, especially towards the end of the game, where humans

and agents reached very similar *Cr* values (near .7). The other kinds of agents invariably ended up with *Cr* values below human performance.

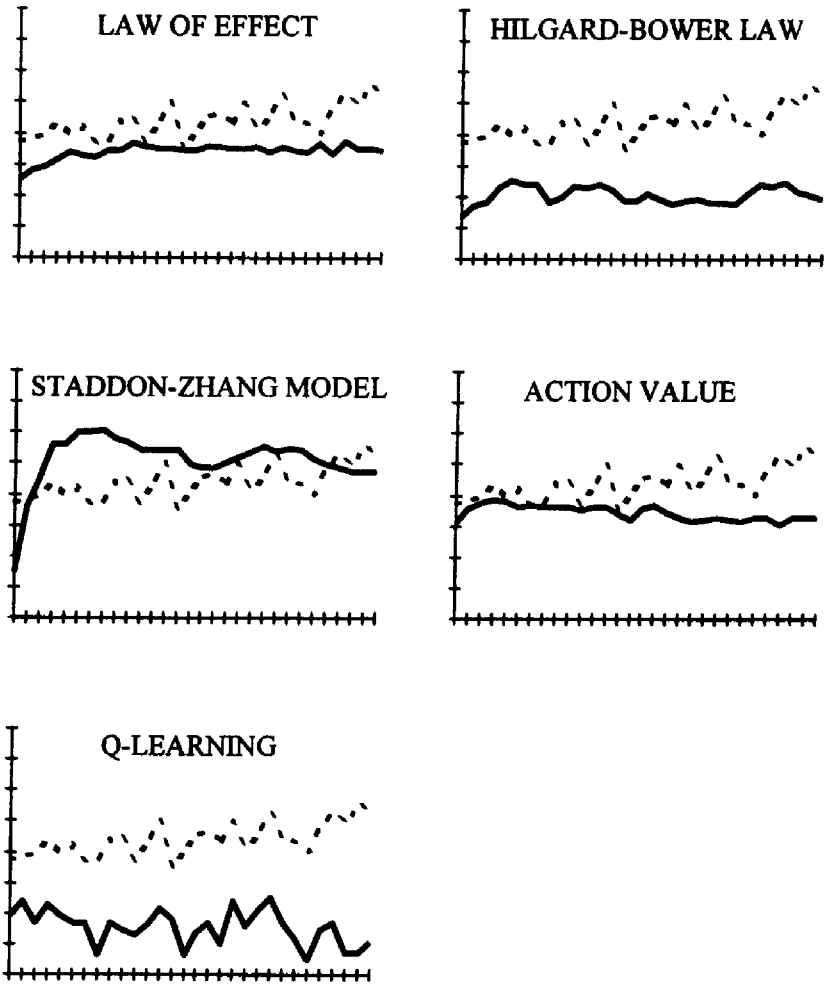

Figure 6. Evolution of the cooperation rate (*Cr*) in human (dotted line) and agent dyads defined according to the five models (continuous lines). In all plots, the ordinate is a linear scale that varies from 0.2 to 1, and represents *Cr*, the ratio between the number of reinforcement behaviors being emitted and the total amount of behaviors being emitted. The abscissa represents the duration of the human experiment (linear scale from 1 to 30 minutes)

The above conclusions must be taken in the context of two caveats.

First, each model relied on a particular set of parameters that determined the agents' behavior and whose impact in a situation of cooperation with humans must be assessed. Second, different results might be obtained with different measures of cooperation. In the present paper, we have limited ourselves to cooperation behaviors. However, if we consider cooperation as a mutual exchange of reinforcements, then measuring the number of behaviors of cooperation emitted by both agents simultaneously would be more appropriate (Delepoulle, Preux, & Darcheville, 1999). In this case, only Q-learning, Staddon-Zhang, Law-of-Effect agents would differ substantially from a random distribution. Moreover, the performance of the Staddon-Zhang agents would remain the closest one to human performance.

GENERAL DISCUSSION AND PERSPECTIVES

We have shown how a selectionist explanation of the emergence of cooperation in human dyads in a minimal social situation is supported by simulations based on reinforcement-learning models. The results obtained with humans may thus be interpreted as resulting from selection by reinforcement, as suggested by Sidowski, Wyckoff, and Tabory (1956) and by Kelley, Thibaut, Radloff, and Mundy (1962). Moreover, in the present paper we showed that such an emergence is more closely modeled by a particular reinforcement-learning scheme, namely, the Staddon-Zhang algorithm. Of course, this algorithm is too simple to capture more complex instances of cooperative behavior or other forms of social behavior. Certain social phenomena may, for example, involve discrimination, a phenomenon that is not readily captured by the Staddon-Zhang algorithm, for this algorithm is, by definition, a free-operant model. Hence, for a more comprehensive understanding of cooperation, other, more complex experimental situations, such as the ones studied by Hake and colleagues, in which case, more complex models would be needed for explanation. But certainly selection by consequences would still constitute a central piece of such models. Far from contradicting a selectionist approach to social behavior, more complex models would be expected to make such an approach more complete.

Our present interest in virtual agents is clearly different from the optimization approach that guides most of engineering applications. Indeed, our aim is to compare selection by consequences in natural and in artificial social situations, rather than obtain optimal solutions to certain practical problems. Our main purpose is to build agents relying on models of behavior selection by consequences. In this sense, the goodness of a model thus depends on how closely its simulation realizations correspond (via either prediction or

postdiction) to the behavior observed in real organisms (humans or animals), even if such simulation and behavior are regarded as being nonoptimal when viewed from an engineering, optimization perspective.

On this basis, the Staddon-Zhang model can be judged as being the best among the other four models in predicting human performance, particularly towards the end of the experiment. This result is likely to be due to the fact that agents designed according to this model can continue responding under intermittent-reinforcement conditions, which involve nonreinforced responses. In fact, these agents keep exploring the state space during the entire simulation. This is the reason why they continued emitting the do-nothing behavior (N) in 30% of the cases in the individual simulations. In this sense, this algorithm captures the idea of behavior selection by consequences more closely than any other model, which may explain why it succeeded in simulating the emergence of cooperation in a minimal social situation. In any case, and beyond particular selectionist models, it is clear that a selectionist behavioral approach allows for the emergence of cooperation under conditions in which agents and human participants alike have no explicit knowledge about the situation. This conclusion is consistent with the ideas about the evolution of cooperation proposed by Axelrod (1984).

REFERENCES

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

Delepoulle, S., Preux, P., & Darcheville, J. C. (1999). Evolution of cooperation within a behavior-based perspective: confronting nature and animats. In *Evolution artificielle 1999*, Dunkerque, Lecture Notes in Computer Sciences. France: Springer-Verlag.

Donahoe, J. W., & Palmer, D. C. (1994). *Learning and complex behavior*. Boston: Allyn & Bacon.

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.

Dougherty, D. M., & Cherek, D. R. (1994). Effect of social context, reinforcer probability, and reinforcer magnitude on humans' choices to compete or not to compete. *Journal of the Experimental Analysis of Behavior, 62*, 133-148.

Hake, D. F., & Vukelich, R. (1972). A classification and review of cooperation procedures. *Journal of the Experimental Analysis of Behavior, 18*, 333-343.

Hake, D. F., & Vukelich, R. (1973). Analysis of the control exerted by a complex cooperation procedure. *Journal of the Experimental Analysis of Behavior, 19*, 3-16.

Hake, D. F., & Olvera, D. R. (1978). Cooperation, competition, and related social phenomena. In A. C. Catania, & T. A. Brigham (Eds), *Handbook of applied behavior analysis* (pp. 208-245). New York: Irvington.

Hake, D. F., Vukelich, R., & Kaplan, S. J. (1973). Audit responses: responses maintained by access to existing self or coactor scores during non-social parallel work, and cooperation procedures. *Journal of the Experimental Analysis of Behavior, 19,* 409-423.

Hake, D. F., Vukelich, R., & Olvera, D. (1975). The measurement of sharing and cooperation as equality effect and some relationship between them. *Journal of the Experimental Analysis of Behavior, 23,* 63-79.

Hilgard, E. R., & Bower, G. H. (1975). *Theories of learning (fourth edition).* Englewood Cliffs, NJ: Prentice-Hall.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: a survey. *Journal of Artificial Intelligence Research, 4,* 237-285.

Kelley, H. H., Thibaut, J. W., Radloff R., & Mundy, D. (1962). The development of cooperation in the "minimal social situation". *Psychological Monographs, 76,* 538.

Olvera, D. R., & Hake, D. F. (1976). Producing a change from competition to sharing: Effects of large and adjusting response requirements. *Journal of the Experimental Analysis of Behavior, 26,* 321-333.

Schmitt, D. R. (1976). Some condition affecting the choice to cooperate or compete. *Journal of the Experimental Analysis of Behavior, 25,* 165-178.

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

Schmitt, D. R. (1986). Competition: Some behavioral issues. *The Behavior Analyst, 9,* 27-34.

Sidowski, J. B. (1957). Reward and punishment in a minimal social situation. *Journal of Experimental Psychology, 55,* 318-326.

Sidowski, J. B., Wyckoff, B., & Tabory, L. (1956). The influence of reinforcement and punishment in a minimal social situation. *Journal of Abnormal Social Psychology, 52,* 115-119.

Skinner, B. F. (1938). *The behavior of organisms.* New-York : Appleton-Century-Crofts.

Skinner, B. F. (1950). Are theories of learning necessary? *Psychological Review, 57,* 193-216.

Skinner, B. F. (1953). *Science and human behavior.* New York: Macmillan.

Staddon, J. E. R. (1983). *Adaptive behavior and learning.* Cambridge: Cambridge University Press.

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. (1988). Learning to predict by the methods of temporal differences, *Machine Learning, 3,* 9-44.

Sutton, R. S. (1998). *Reinforcement Learning.* Cambridge, MA: MIT Press.

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.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative process in animal. *Psychology Monographs, 2*, 8.

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies*. New York: MacMillan.

Watkins C. J. C. H., & Dayan, P. (1992). Q-learning. Technical note. *Machine Learning, 8*, 279.