Cooperation is not always so simple to learn

M. Mailliard, F. Amblard, C. Sibertin-Blanc

IRIT – Université de Toulouse 1
21, allées de Brienne, 31 042 Toulouse Cedex – France
mmaillia@univ-tlse1.fr, sibertin@univ-tlse1.fr, famblard@univ-tlse1.fr

Abstract. In this paper, we propose to study the influence of different learning mechanisms of social behaviours on a given multi-agent model (Sibertin-Blanc et al., 2005). The studied model has been constructed from a formalization of the organized action theory (Crozier and Friedberg, 1977) and is based on the modelling of control and dependency relationships between resources and actors. The proposed learning mechanisms cover different possible implementations of the classifiers systems on this model. In order to compare our results with existing ones in a classical framework, we restrain here the study to cases corresponding to the prisoner's dilemma framework. The obtained results exhibit a variability about convergence times as well as emergent social behaviours depending on the implementation choice of classifiers systems and on their parameters. We conclude by analysing the sources of this variability and by giving perspectives about the use of such a model in broader cases.

1 Introduction

The way social actions are coordinated by and among social actors has been a source of inspiration for different theories in very different domains. For instance game theory in economics, their application in ecology (Dugatkin, 1984) and other related theories either in sociology or even in psychology. In a larger perspective, we choose as a research project to investigate the sociological theory of organized action proposed by Crozier and Friedberg (1977), on the one hand to improve this discursive theory by proposing a formalisation of it and on the other hand to apply this theory to model different social phenomenon appearing in organizational contexts. The work conducted on this project resulted in a proposed formalization of this theory (Sibertin-Blanc et al., 2005), a meta-model, that we expose briefly in the first section.

Taken for granted this model, we are then searching to improve it by including social learning mechanisms in order to take into account the strategic rationality of the actors. We then focus on classifiers systems as a way of implementing social learning. Such a learning mechanism, even simple, leads to different implementation choices in order to adapt it to the existing model. Different possibilities being possible and realistic, we decided to study those alternatives in order to make up our mind. We present those alternatives in the second section as well as their sociological interpretations in the frame of the organized action theory.

Given those learning mechanisms, we are searching to understand in each cases, which collective strategies could emerge and why. We then proceed to experimentations on each alternative (section 4). In order to make things understandable at a first attempt, we choose to focus on the particular framework of prisonner's dilemma, which is reproducible by our model. We then choose a two players game and we make vary the parameters corresponding to the share of resources among the actors.

The experimentations exhibit a variability of collective behaviours depending on the chosen parameters as well as a phase transition at the tipping-point corresponding to the individual transition from dependency on the resources and control of those resources. Results are given section 5. In section 6, we provide conclusions concerning the comparisons of the different learning mechanisms as well as the observed phase transition. We conclude by giving some of the steps following this work.

2 Formalization of the Organized Action sociological theory

A formalization of the Sociology of the Organized Action (SOA) leads to consider that constitutive elements of a social system are of the three different types shown in Fig. 1: the Actor, the Relation and the Resource.

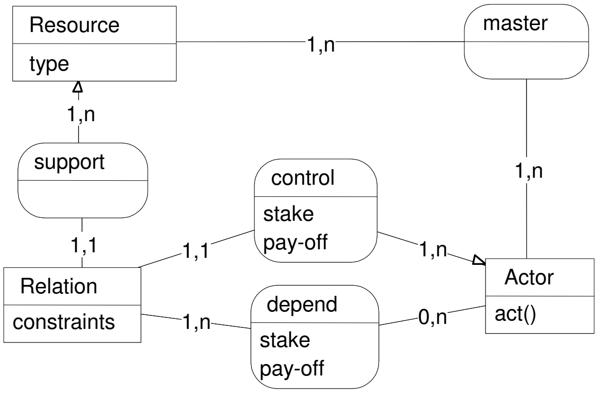

Fig. 1. Model of the structure of a social system in the frame of the SOA (using the Entity/Association formalism).

A *Resource* can be the support of one or more *Relations* associated to *Actors* who are linked to it, either because they *control* the Relation or they *depend* on it. Each actor puts *stakes* and receives in return a *pay-off* for each one of the Relations he is implied in. The actor who *masters* a Resource (by the mean of a Relation he controls) decides of the distribution of the pay-offs among the actors who depend on this Relation. The Resources of a CAS are the necessary elements for the organized action, their availability being required in order to make some action. Every Resource is *mastered* by one or more Actors who decide about its availability and therefore influence the action capability of the Actors who need it. Each Resource leads to the creation of one or several Relations. A Relation is unbalanced as a unique Actor *controls* this Relation while other Actors depend on this Relation as they need this Resource to achieve their goals. Every *Actor* masters one or more Resources and then possesses some freedom to act that he exerts by means of the Resources he controls.

Each Actor distributes his *stakes* on each one of the Relations he participates to, either by controlling them or depending on them, depending on the importance of the Resource in regards to his objectives. The Actor controlling a Relation is the one who determines the *pay-off* other actors received from the Resource. The pay-off corresponds to the quality of the Resource availability; more or better the usability of the Resource by an Actor, higher his pay-off for this Relation.

The distribution of pay-offs and stakes on numerical scales enables, applying simple operations, to aggregate those values in synthetic and significant values. One can graduate the stakes on a scale between 0 and +10, and the pay-offs with the correspondence –10 to +10. As evidence, these numerical values just enable to perform comparison among them. To do so, we have to normalize the sum of the actors' stakes and then attribute the same amount of stakes to each actor for him to distribute on the relations he participates to. This normalization comes down to grant the same investment to each actor, the same possibility of personal implication in the social relations game.

A particularly significant value is, for each actor, the sum on the whole set of Relation he is involved in, of a combination between his stake and the resulting pay-off he receive. We name this value the actor's *satisfaction* (rather than utility because it is more linked to a bounded rationality context). It expresses the possibility for an actor to access to the resources he needs in order to achieve his objectives, and then the means available for him to achieve these objectives. A linear version consists in considering the sum, on every relation he is involved in, of the stake by the pay-off:

$$Satis(a) = \sum_{r/\ a\ participates\ to\ r} stake(a, r) * pay\text{-}off(a, r)$$

To obtain or preserve a high level for this satisfaction is a meta-objective for every actor, as this level determines his possibility to achieve his concrete objectives. The strategic characteristic of an actor's behaviour leads him, by definition, to aim and achieve his objectives and then to obtain an acceptable value (if not the optimum) for his satisfaction, that becomes the criterion for learning mechanisms we expose in the following section.

3. The learning mechanisms implemented

After the claim by Rosaria Conte (2001) among others for integrating so-called « intelligent » social processes in the agents when doing agent-based social simulations, several papers (Conte & Paolucci, 2001 ; Flache & Macy, 2002 ; Takadama 2003) proposed and implemented some learning algorithms to be used by agents.

In this paper we propose to explore two models for social learning using the classifiers mechanism (Holland et al., 2000) for the selection of the action; it is based upon the learning of behavioural rules by test-errors and reinforcement of the rules depending on the results they produce. Recent works about reinforcement learning models exhibit that a reduce set of parameters and hypothesis may cover hidden important theoretical assumptions (Macy & Flache, 2002). Therefore we decide to follow the Axelrod's (1997) famous maxim : "Keep it simple, stupid". Thus, each model is a naive answer aiming at validating or not a less naive question on social learning.

```
Do
    For each actor a
      Satisfaction_computing()
          For each relationship r activated by a at time t
                Retribution_Process(elected_rule)
                Oblivion_Process(r)
      For each relationship r controlled and activate by a,
                [Mr] ← Matching (r)
                er ← Electing_Process (Mr)
                if er = null, then er ← Covering_Process(r)
      For each actor a
            Act(a, er)
  Until the end of the simulation
```

Table 1 Pseudo code of the Classifier System algorithm used in the social learning models

Following the rationality hypothesis implied by the Sociology of the Organized Action, the two models are based on the standard three phases cycle: perception of its own state and of the environment; selection of an action to perform, according to its expected effect on the gap between the current and the goal state, execution of this action.

Each model is a kind of Learning Classifier System (LCS) without genetic algorithms nor bucket brigad retribution process (Table 1). The main processes involved, namely retribution, oblivion and matching are respectively governed by three parameters: *reward* acts for the positive or negative reinforcement of the rules depending of $\Delta(\textit{satisfaction})$ sign; *oblivion* is a factor of *reward* used to weakened the strength part of each rules and erase useless ones; *dmin* enables an agent to match a perceive situation with existing situations of the learned rules. The election process selects the matching rule with the highest strength and if there is no elected rule the covering process generates a random one.

The two models differ essentially in the way satisfaction is computed and involved in the reinforcement procedure of the learned rules.

3.1. 1st solution: ***Independent Multi-Satisfaction CS***

How in a social environment, a learning actor must adapt itself from the others behaviours? We want an actor able to adapt to different actors but also to different kinds of relationship (friendship, employement, family…). This disctinction in relationships is embedded in the meta-model expressiveness as the concept of *relation*. Therefore the base hypothesis underlying this first model is that an agent adapts itself within each of its *relation* so that it can adapt itself in various kinds of relationships independently.

We thus propose to associate a satisfaction to each controlled relation and to retribute only the associated elected rules. Each actor *a*'s *satisfaction* associated to a relation *r* is expressed as: $\text{Satisfaction}_{\text{a,r}} = \text{stake}_{\text{a,r}} * \text{pay-off}_{\text{a, r}}$. We will refer to this kind of CS as Independent Multi-Satisfaction CS or IMSCS.

3.2. 2nd Solution: *Global Satisfaction CS*

As exprimed by Molm, many social scientists have stressed out the point that satisfaction may be specific (IMSCS) or global. We would now address this second point by a question pointing to Molm's definition of satisfaction: how can we make global "*cognitive evaluations in which actors compare actual to expected outcomes*"?

We propose a raw answer by aggregating each specific satisfaction as defined in the IMSCS model and by summing up them into a global satisfaction. In such a way an agent will reinforce all the elected rules in the same way. In this Global Satisfaction CS (GSCS), the satisfaction is given by: $\text{Satisfaction}_{\text{a}} = \sum_{\text{r/ a participates to r}} \text{stake}_{\text{a,r}} * \text{pay-off}_{\text{a,r}}$

4. Experimentations conducted

In order to validate our model we propose to used a cross validation as proposed by Takadama in regard to a famous well-known by game theorists: the prisonner's dilemma (PD). Althougth Takadama work has been a strong and rich influence in conducting our researches, our experimental protocol is not quite the same. First, we proceed using an exhaustive space parameters' exploration of the learning models (oblivion/reward, d_{min}) and of the social model (stakes). This exploration goes beyond the constraints of the prisonner's dilemma and enables to situate the results in a wider area. Secondly, the agent's representation is the same for every models. Finally, the criteria used to validate the models is not the one of a perfect rationality but, with respect to Simon (1996), a bounded rationality, that implies imperfect actors but not foolish one.

4.1. The Prisonner's Dilemma

Description of the prisonner dilemma game

The prisonner's dilemma was first proposed by two mathematicians Merrill Flood and Melvin Dresher in 1950. It is exposed as a game where two players have the choice between two actions: cooperate or defect. Players earn pay-offs depending on the choices of the both players, as shown in Table 2.. That is if the two players cooperate (CC) they will receive the reward for the cooperation (R); if both defect they will be punished for the defection (P); and if one cooperates whilst the other defects, he is the sucker (S) and the other will ear the retribution of his temptation (T).

	c	d
c	R, R	S, T
d	T, S	P, P

Table 2. The pay-offs matrix for the prisonner dilemma game

The dilemma is constrained by the fact that temptation is more profitable than mutual cooperation, that pays more than punishment, that is more valuable than the sucker: $T > R > P > S$. Therefore the dilemma is shown when an actor is tempted to defect and he infers that other behaviour could be the same as him, so he would prefer to cooperate but what if the other defects. An other inequation, $2\,R > T + S$, encourages cooperation by giving a prior account to mutual interest than to selfish one.

The classical PD game is of minor interest compared to its iterated version where each player can potentially apply different actions over time and where the pay-offs are summed up. The iterated version of the PD has been widely explored and exposed (Hoffman 2000, Delahaye 1992, Macy & Flache 2002) since Axelrod's works (1984). In his tournaments, Axelrod has found many interesting emergent strategies those most famous is the Tit-forTat one: a simple, robust and ethic strategy.

Adapting the sociology of organized action formalization to PD game constraints

The SOA formalization expressivity does not directly match the PD game. So we will pesent here how we make a projection from our model to a PD game context.

Let be two actors, *a1* and *a2*, participating in two relations, *r1* and *r2*, such that each actor controls one relation. Let the sum of the stakes for each actor be normalized to 10. Let be $s_{r,a}$ and $p_{r,a} \in [0;10]$ respectively the stake and the pay-offs of an actor *a* for a relation *r*. Let be *give* and *take* the possible actions each controler can exert on a relation.

We now define the effect of an action *action* applied by the controler *c* of a relation *r* as $effect_r(action)=\ \{\Delta p_{r,c}, \Delta p_{r,d}\}$ such that $\Delta p_{r,c}$ and $\Delta p_{r,d}$ are respectively the pay-off increments of the controler *c* and the actor dependant *d* of the relation *r*. Let be $effect(give)=\ effect^{-1}(take)=\{-1,1\}$.

		a1	
		give	*take*
a2	*give*	$R1 = s_{r2,a1} - s_{r1,a1}$ $R2 = s_{r1,a2} - s_{r2,a2}$	$T1 = s_{r1,a1} + s_{r2,a1}$ $S1 = - s_{r1,a2} - s_{r2,a2}$
	take	$S1 = -s_{r1,a1} - s_{r2,a1}$ $T2 = s_{r2,a2}+ s_{r1,a2}$	$P1 = s_{r1,a1} - s_{r2,a1}$ $P2 = s_{r2,a2} - s_{r1,a2}$

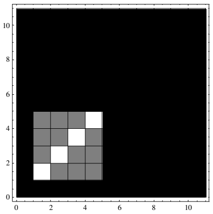

Fig.2 Pay-off matrix for the specified CAS; Graphics of the fully and quasi-satisfied PD constraints (X-axis: $s_{r1,a1}$, Y-axis: $s_{r2,a2}$). We do not have to represent $s_{r2,a1}$ nor $s_{r1,a2}$ because of the stake normalization. White squares represent the fully satisfied PD constraints while grey ones represent the quasi-satisfied PD constraints.

In Fig. 2, the matrix gives the pay-offs for the defined CAS. At the difference of the classical PD pay-offs we obtain a potentially different pay-off for each actor. The graphic in Fig. 2 gives, in white, the cases where the PD game constraints are fully satisfied and, in grey, the cases where they are quasi-satisfied (R1 > S1 > T1 > P1, R2 > S2 > T2 > P2, R1+R2 > T1+S2 and R1+R2 > T2+S1).

4.2. Experimental design

The simulations where products with the same experimental design for the both models, that is the IMSCS and the GSCS. For each one, we have conducted a systematic exploration for two kind of parameters which are present in both models.

The first set of parameters directly concerns the sociological model. It is composed of the stake of each actor for the relation he controls. We have not explored the stake of each actor for the relation he is dependant because the stakes normalization directly constraints the value of the later from those of the former. In order to accelerate the computation of the large amount of runs we have take into account of the symetric nature of the stake matrix (Fig.3) we want explore. Thus, we only show the computed part of the symetric matrix. This matrix permits us to present many observations for all the possible integer values for all the implied stakes with respect to the previously mentionned optimisation. Values in the matrix are given by greyscale enabling an observer to quicly acquire and compare all the avaiable information: 11*(11+1)/2 datas for each matrix. We sometimes apply a contour filter.

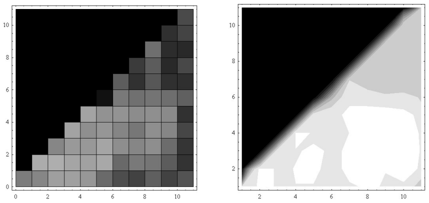

Fig.3 Examples of data observations in the stake matrix with and without contour filter

The second set of parameters concerns the learning models. We explore the *dmin* and the *oblivion/reward ratio* as follow. The dmin is an essential parameter permitting to explore and to learn from situations in the space of phases[1]. The greater is dmin the less thick is the exploration. The possible values we have choosen to explore dmin are graduated on a logarithmic scale from 0, 2^0,… to 2^5. As an illustration the 2^5 upper bound value leads the agents to consider each position in the space of phases as belonging to the same situation; whilst the opposite bound value, 0, will lead to considere each position as a new situation and will thus multiply the learned rules. The oblivion/reward ratio is also essential because it permits to renew the rules population those situation part is matching the current situation. A high ratio value will conduct to a quick renewing of the population, and at the opposite a low ratio will slow down the adaptation of the agent. The ratio value belong to [0;1] and is incremented with 0.2 step, that is we have 6 values. The reward is fixed at 5.

We have produced 50 runs for each parameter quadruplet {stake_of_r1_controler, stake_of_r2_controler, dmin, oblivion/reward}. For each model and for each parameter quadruplet we have observed the following values: the mean and the standard deviation for the convergence within the limits of 200 steps, and the occurences of a convergence toward a CC (give_give), CD/DC(give_take/take_give), DD(take_take) situation.

Finally, all the simulations have been implemented under Java, and most of the data analysis has been made under Mathematica.

5.Results

5.1. Results for the Independent Multi-Satisfaction Classifier System

The Fig. 4 shows the results for the ICSCS model. The left and right matrixes respectively present the observations for the mean and the standard deviation of the convergence.

How to read this five dimensions representation ?

- Each main matrix contains the previously introduced pay-offs matrixes.
- The X-axis of the main matrice represents the dmin parameter while the Y-axis represent the oblivion/reward ratio.
- The greyscale semantic for the mean convergence is *dark grey* for a quick mean convergence to white for the upper limits of 200.
- The greyscale convergence for the standart deviation is *black* for a zero deviation and white for deviation upper 100.

[1] More precisely it is a projection of the space of phases on the individual stakes and pay-offs dimensions.

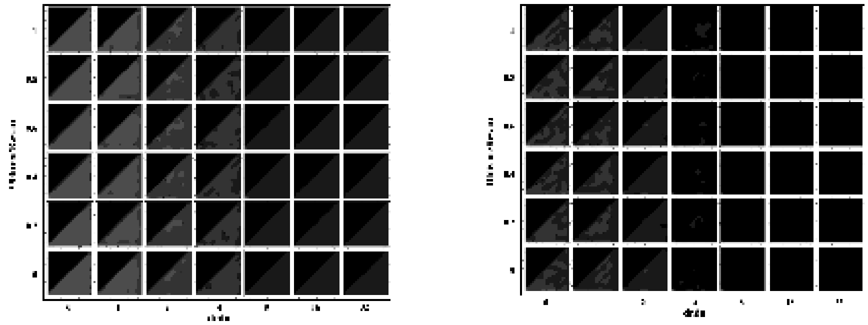

Fig.4 Observations of the mean and standard deviation of convergence.

We can observe that the mean convergence is decreasing with the dmin parameter while the oblivion/reward ratio seems to have no effect on it. The convergence is generally quite quick. In every cases the standard deviation is near or equal to zero.

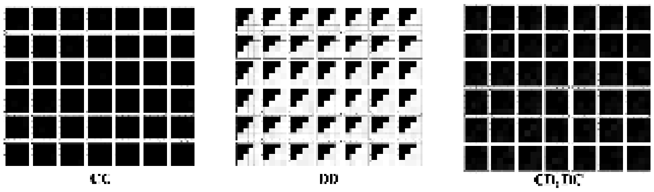

Fig.5 Observation of the occurrences of learning the different action pairs.

The Fig.5 present the observations of the occurences of the caracteristic situations resulting in the social learning of the action pairs CC, DC and CD/DC. We only shows the part of the pay-offs matrixes which validates the fully and quasi-satisfied PD constraints as shown in Fig.2 where $s_{r1,a1}$, $s_{r2,a2} \in [1;5]$. The greyscale semantic is given by the application [black; white]→[0 occurrence; 50 occurences]. We can clearly and only observe that every cases give place to a DD learning for every runs.

5.2. Results for the Global Satisfaction Classifier System

The GSCS seems to give a largest variety of results than the IMSCS. As we can observe in the mean matrix, on the left of Fig. 6, that the dmin parameter globally decrease the time to converge as it increase, and that the oblivion/reward ratio also speed up the convergence as it is valued between 0.2 and 0.6. It also appear on the rigth matrix of standard deviation that the dmin decrease have a clear tendance to increase deviation. So, globally and for a given parameter quadruplet, agents take many ways for co-adaptate themselves, sometimes it is quick and sometimes it is not.
An other observation present in both the mean and the standard deviation matrixes is a distinct phase transition which strangely appears in the area where the PD game constraints are fully or quasi-satisfied. The convergence in this region is clearly more slow and more various than in other ones.

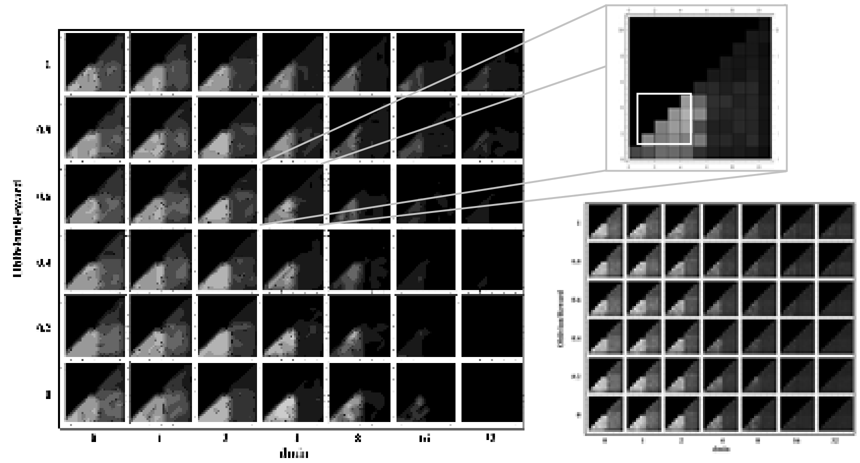

Fig. 6 Observations of the mean and standard deviation of convergence. A clear transition phase appears in the area of fully and quasi-satisfied PD game constraints.

By observing the occurences of action pairs in Fig. 7 it globally appears that DD learning is more frequent. If we make a more precise and specific analysis we can notice that: the CC pairs emerge more frequently as the dmin value is strictly upper 8 and the oblivion/reward rate increases; the DD pairs emerge more frequently as the dmin value is strictly upper 1 and the oblivion/reward rate decreases; the CD emerges more frequently as the dmin value is under 16 and the oblivion/reward rate increases.

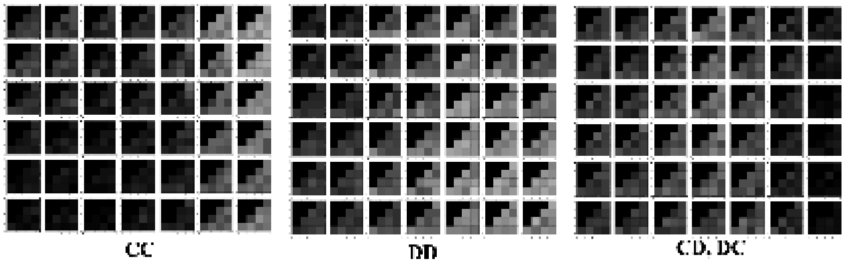

Fig. 7 Observations of the occurrences of learning the different action pairs

The final global observation is that CC is the globally less easy action pair to learn. Our exhaustive exploration permits to chose the adequate values of dmin and the oblivion/reward ratio to palliate the problem. But the choice of these values has a high cost in the reactivity to adapt to perturbation, because they imply the stagnation of the population rules. Althougth in the proposed PD game there are no such perturbation and the solution is acceptable.

6.Discussion

6.1.Comparison of IMSCS and GSCS

The two proposed learning systems are naive answers to complex questions. We still not have the rigth answer to the generally addressed question of what is satisfaction and what is its role in agent learning processes. Althougth we have proposed to validate the rationality of our imperfect agent models by confronting them in the context of the prisonner dilemma game. We also have systematically explorated the space of phases and thus the emergent potentialities of our models.

With regard to the bounded rationality and only in this PD context, it is clear that the IMSCS exhibits a quite straightforward irrationality in the way he adapts himself only to the relations he controls. And of course it is not validated because he can only learn to defect. But it could be more pertinent, and thus reused, in other contexts.

The GSCS exhibits good qualities to be a serious candidate to the position of bounded rational strategic actor in a PD game. It can learn all the kind of action pairs. Its main problem was the real difficulties to learn cooperation as quickly as the other possibilities. But we can partially correct some of its imperfection by selecting the good values for different internal parameters. So this model is validated.

6.2.Explanation of the observed phase transition

The observed threshold correponds to the shift from a mutual dependence situation, where each actor need the usage of a non controlled ressource, to an asymetric dependence or an independant situationAs an evidence, without motivation to cooperate agents prefer an other, and thus quickest way to act. For example the game with the same stake for each relation and each actor will lead to T1 > R1 = P1 > S1 and T2 > R2 = P2 >S2 which present a Nash equilibrum for D. Without giving all the details, it will quickly converge to learn DD, and sometimes, if dmin is high enougth, to learn CC. If sr1,a1 > 5 > sr2,a2, the PD pay-offs constraints are respected for a_2 but the fact that T1 > P1 > R1 > S1 is a real motivation for a_1 to defect. It is the same motivation for both agent if sr1,a1 > 5 > sr2,a2.

7.Conclusion

As an evidence, the presented case is constraining considering the learning mechanisms that could be proposed. Following works by increasing the number of actors or by using different oragnisation of the relations between actors and resources could lead to suggest other learning mechanisms. But in this case, as a result of our study, the second mechanism is more adequate to our modelling purposes.

We have also to mention that the proposed model, as a formalisation of the sociological theory of organized action, has a far broader spectrum of application than the one presented in the paper. Among others, we applied or derived this meta-model on the study of the emergence of territorial coalitions (Mailliard et al., 2005), to classical cases from the strategic analysis literature as the Trouville case (Mailliard et al., 2003).

Moreover, this interdisciplinary work even presented as the use of computer sciences as tools for sociological theories, benefits also to computer sciences as a source of inspiration in order to propose original coordination mechanisms among computational agents (Sibertin-Blanc et al., 2005).
d'utiliser ce méta-modèle comme modèle de coordination pour les systèmes multi-agents (Sibertin-Blanc et al., 2005).

References

Axelrod, R.: "The Complexity of Cooperation: Agent Based Models of Complexity and Cooperation", Princeton University Press, (1997).

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

Conte, R., Paolucci, M.: Intelligent Social Learning. Journal of Artificial Societies and Social Simulation JASSS vol. 4, no. 1, (2001).

Crozier, M., Friedberg, E. : L'acteur et le système : Les contraintes de l'action collective. Seuil (1977).

Delahaye, J. P., L'altruisme récompensé ? *Pour La Science (French Edition of Scientific American)*, 181:150-156, (1992).

Dugatkin, L.A.: Cooperation among Animals: An Evolutionary Perspective. Oxford University Press, (1997).

Flache A., Macy, M.W.: Stochastic colusion and the power law learning. Journal of Conflict Resolution, (2002).

Hoffmann, R.: Twenty Years on: The Evolution of Cooperation Revisited, Journal of Artificial Societies and Social Simulation (JASSS) vol. 3, no. 2, (2000).

Holland, J., Booker, L.B., Colombetti, M., Dorigo, M., Godberg, D.E., Forrest, S., Riolo, R., Smith, R.E., Lanzi, P.L., Soltzmann, W., Wilson, S.W.: What Is a Learning Classifier System? LCS'99, LNAI 1813, 3-32 (2000).

Macy, M.W., Flache. A.. Learning Dynamics in Social Dilemmas. Proceedings of the National Academy of Sciences U.S.A. May 14;99(10):7229-36,(2002).

Mailliard,M.,Amblard, F., Sibertin-Blanc C.: Modélisation multi-agents pour la formalisation de théories sociologiques: Le cas de la sociologie de l'action organisée appliquée à l'étude de la dynamique du pays Quercy-Rouergue. In Proceedings of the SMAGET Conference, Bourg St. Maurice, France, (2005).

Mailliard, M., Audras, S., Casula, M. : Multi Agents Systems based on Classifiers for the Simulation of Concrete Action Systems. In Proceedings of the 1st EUropean Workshop on Multi-Agent Systems (EUMAS), Oxford University, (2003).

Molm, L.: Affect and social exchange: satisfaction in power-dependence relations. American Sociological Review, vol. 56, (1991).

Sibertin-Blanc, C., Amblard, F., Mailliard, M.: A coordination framework based on the Sociology of the Organized Action, In Proceedings of the From Organization to Organization Oriented Programming in MAS, AAMAS, Utrecht University, (2005).

Simon, H.: The sciences of the artificial, MIT Press, 3rd edition (1996).

Takadama, K., and al.: Cross-Element Validation in Multiagent-based Simulation: Switching Learning Mechanisms in Agents. Journal of Artificial Societies and Social Simulation vol. 6, no. 4, (2003).