José Castro Caldas and Helder Coelho (1999)

The Origin of Institutions: socio-economic processes, choice, norms and conventions

Journal of Artificial Societies and Social Simulation vol. 2, no. 2,
<http://jasss.soc.surrey.ac.uk/2/2/1.html>

To cite articles published in the *Journal of Artificial Societies and Social Simulation*, please reference the above information and include paragraph numbers if necessary

Received: 2-Feb-99 Accepted: 10-Mar-99 Published: 14-Apr-99

Abstract

Institutions, the way they are related to the behaviour of the agents and to the aggregated performance of socio-economic systems, are the topic addressed by this essay. The research is based on a particular concept of a bounded rational agent living in society and by a population based simulation model that describes the processes of social learning. From simple co-ordination problems, where conventions spontaneously emerge, to situations of choice over alternative constitutional rules, simulation was used as a means to test the consistency and extract the implications of the models. Institutions, as solutions to recurring problems of social interaction, are both results and preconditions for social life, unintended outcomes and human devised constraints. In an evolutionary setting no support is found for the deep rooted beliefs about the 'naturally' beneficial outcomes generated by 'invisible-hand' processes or by any alternative Hobbesian meta-agency.

Keywords:
Institutional Economics, Agent Modelling, Socio-economic Simulation, Evolutionary Algorithms

Introduction

1.1

An old enigma has not been answered in a satisfactory way: human beings, even though competing for scarce resources, are unable to live without one another; how is it that more or less stable patterns of relations are established that ensure the conditions for social life? Nature seems indeed to have played "a cruel trick on our species - we cannot survive alone, yet unlike social insects we are not genetically hardwired for co-operation" (Macy, 1998).

1.2

The founding fathers of economics, first of all Adam Smith, had an answer. At a time when it was generally believed that the welfare of nations rested on the wisdom and benevolence of their rulers, Smith argued instead that economic order and welfare existed in spite of the rulers, and should be

understood as an unintended result of the actions of a multitude of individuals who pursue their own interests. Walras transported this intuition to present day Economics by means of his general equilibrium theory. This model, in the sophisticated mathematical form given to it by Arrow and Debreu, shows that in order to rigorously prove Smith's two main conclusions: (a) the existence of a state of general agreement over a set of relative prices, and (b) the 'good' properties of this state of affairs, a number of (hard to swallow) assumptions have to be imposed on the world. In fact, it is common knowledge among economists that the existence of a market order with efficiency properties can only be proven if perfect rationality, competition, non-increasing returns, and the absence of public goods and external effects are assumed.

1.3

We would not have to worry much about the realism of those assumptions if there were signs encouraging us to believe that in the end the model will turn out to be robust, in the sense that the main conclusions will still prevail when these 'simplifications' are relaxed. While many still believe that this may be the case, and keep working on reformulating the general equilibrium model, a growing minority feels that the difficulties that this theory is facing simply show that the explanation for the enigma of society has to be looked for elsewhere. Given the conditions in the real word, the economic order may depend much more than the general equilibrium theorists assumed on the institutional frame of the interaction; and, this may be taken as an invitation to redirect the research efforts of the economists towards alternative accounts of the social order and human action.

1.4

In recent years the topic of institutions has re-emerged in economics, giving rise to a vast debate and to a movement for an Institutional Economics, that permeates different traditions and economic paradigms[1]. The primary aim of this paper is to contribute to this discussion from a perspective that combines ideas, methods and formalisms from Economics and AI (Caldas and Coelho, 1994). Since there is a striking parallelism between the problems that Institutional Economics, other Social Sciences and Distributed AI (DAI) are faced with, researchers in those fields will easily recognise the issues we are dealing with.

1.5

What is the origin and role of conventions and shared rules, and how are they related to the performance of socio-economic systems? What is their relationship to the behaviour of the agents? These are the questions addressed in this paper. The research is supported by a simulation tool that was built on a simple model of an economic agent that is purposeful but limited in perception and cognition. Emphasis is given to system mediated interaction, that is, to a particular type of economic environment that excludes direct communication between agents.

The unidimensional mind

2.1

Herbert Simon's critique of the rational choice paradigm and his own concept of bounded rationality had a huge impact both in Economics and AI. However, Simon's scenario in his seminal papers (1955, 1956) was one of a single agent in interaction with the world, or at best two agents over a chessboard (the typical closed world assumption). The relevant aspects of decision making in social settings were not taken into account. It is useful to revisit Simon's critique and model placing now the decision making agent within a social environment, but before that an even more fundamental problem with the model of economic man must be considered.

2.2

Long ago Edgeworth (1881: 16) wrote: "The first principle of Economics is that every agent is

actuated solely by self-interest". This is still the principle on which the game theoretical/economic standard model is founded. The problem is that this 'first principle', which may seem crystal clear at first sight, soon becomes confusing as the question of what might be considered to be the interest of the agent is posed: his own well being? the well being of his family, of his neighbours, of his country? In fact, if the scope of self-interest is indefinitely extended, any act, no matter how 'altruistic' it may seem, can be interpreted as self-interested (or even egoistic). There follows a game of words that leads nowhere, turning Edgeworth's sentence into a mere tautology. Jevons (1871: 25), however, had clarified exactly what the marginalists (including Edgeworth) had in mind:

> As it seems to me, the feelings of which a man is capable are of various grades. He is always subject to mere physical pleasure or pain [...]. He is capable also of mental and moral feelings of several degrees of elevation. A higher motive may rightly overbalance all considerations belonging even to the next lower range of feelings; but so long as the higher motive does not intervene, it is surely both desirable and right that the lower motives should be balanced against each other[...]. Motives and feelings are certainly of the same kind to the extent that we are able to weigh them against each other; but they are, nevertheless, almost incomparable in power and authority.
>
> My present purpose is accomplished in pointing out this hierarchy of feeling, and assigning a proper place to the pleasures and pains with which the Economist deals. It is the lowest rank of feeling which we here treat. [...] Each labourer, in the absence of other motives, is supposed to devote his energy to the accumulation of wealth. A higher calculus of moral right and wrong would be needed to show how he may best employ that wealth for the good of others as well as himself. But when that higher calculus gives no prohibition, we need the lower calculus to gain us the utmost good in matters of moral indifference.

2.3

This long quotation could not be avoided because it makes two points very clearly:

1. Economics was supposed to deal with the "lowest rank of feeling" under the assumption of the absence of motives arising from any "higher ranks";
2. The 'feelings' pertaining to different levels are incommensurable, they are "almost incomparable in power and authority".

These two points have important implications. First: since the absence of motives arising from any "higher rank" can only make sense in a situation of interaction where actions that have positive consequences for an agent do not affect all the others, the remaining situations (including therefore a large section of the subject matter of economics and game theory) are out of the scope of the economic-man model. When external effects are present, there are "no matters of moral indifference". Second: in no way can the 'hierarchy of feelings' be aggregated in a single (context independent) utility function. Any "calculus of moral right and wrong" must involve not only individual values, but the context dependent measure in which those values are shared and respected within the group. Therefore, Edgeworth's unidimensional mind concept of self-interested action does not fit in environments other than the typical economic situation of anonymous interaction with productivity related rewards.

2.4

But the unidimensional mind is also present, although differently, in alternative accounts of human action. In what might be called a standard functionalist sociological explanation for the fact that individuals tend to behave in accordance with social norms, the emphasis is on socialisation, "a process in which, through (positive and negative) sanctions imposed by their social environment, individuals come to abide by norms", and that leads to internalisation, "according to which a

person's willingness to abide by norms becomes independent of external sanctions and, instead, becomes part of a person's character" (Vanberg, 1994: 14).

2.5

There is at least one problem with the functionalist explanation: "By invoking at the same time, through the concept of sanctions, that people respond to incentives and, through the notion of internalisation, that their rule compliance is unresponsive to incentives ... [it] seems to be based on two incompatible conceptions" (Vanberg, 1994: 14). In fact, it is hard to accept that the "willingness to abide by norms" is second nature. Should we believe that those who comply with social norms will continue to do so even when they are confronted with a situation where 'mutant' behaviour is rewarded and those norms become useless?

The hierarchy of feeling

3.1

If Economics is to deal with interactions within society where strong external effects are present, the model of man will have to be reconsidered. We are no longer dealing with "the lowest rank of feeling". In such circumstances to comply with a shared rule may be a motive behind choice. The agents may be endowed with a moral disposition (Vanberg, 1994) that drives them to behave in accordance with the rules that are believed to sustain the group's existence. However, even if we accept that a moral disposition has to be taken into account, this propensity cannot be viewed as a fixed parameter in the agent's model. The moral disposition is not imprinted once and for all, and in the same degree to all agents, by genetics or culture. Since a shared rule can only produce the expected benefits if it is generally abided by, it may become pointless not to violate it when most others do. The moral disposition is therefore a variable in two senses: it varies from individual to individual, and it tends to be strengthened with rule compliance, and weakened with the spreading of deviant behaviour.

3.2

Two ideas are combined here: (a) The existence of a 'hierarchy of feeling' encompassing normative obligations; (b) The dependence of the moral disposition on the level of compliance with shared rules within the group. The first idea has been developed by different authors. Margolis (1991) presented a model with individuals endowed with two utility functions: purely individual S preferences and purely social G preferences. Buchanan and Vanberg (Vanberg,1994 :21) speak of a distinction between action interests and constitutional interests. The action interest concerns personal situational choices within a set of alternatives. The constitutional interest is related to shared rules and might be defined as the individual's interest "in seeing a certain rule implemented in a social community within which he operates" (Vanberg, 1994 :21). The second idea of a relation between the moral force of shared rules and the degree of compliance within the group was developed by Akerlof (1980) and by Sugden (1986). In the DAI literature, similar concepts can be found in Conte and Castelfranchi (1995).

Solutions to the problem of collective action

4.1

Game theorists and other social researchers have invested a huge effort in trying to show that a social order might spontaneously come into existence and be reproduced without the enforcement of norms by a coercive agency of some kind. In some contexts, related to co-ordination, their arguments are rather convincing. Conventions may emerge as a non-intended outcome of repeated interaction. They have also shown that co-operation is possible in indefinitely repeated Prisoner's Dilemma (PD) games. However, this result is much harder to obtain for N-person games. The possibility of an anarchic order therefore remains open for speculation. It is not completely ruled

out, whether we approach it in game theory terms, or from the historic and anthropological record, but it is far from having been proven convincingly. The Hobbesian solution is much more familiar to us.

4.2

Hobbes's argument on the need for a social contract and an enforcing sovereign power has been translated to modern terminology by game theorists: "the Hobbesian argument essentially turns on the claim that the problem of political obligation can only be solved by the creation of a co-operative political game, instead of the non-co-operative game played in the state of nature" (Heap, 1992 :203). Co-operative games are based on pre-play negotiation and binding agreements. In these terms the social contract would be the result of pre-play negotiation and the presence of Hobbes's sovereign the condition to make this contract binding to all.

4.3

In a Hobbesian vein Edmund Burke (quoted in Vanberg, 1994 :41) moralised: "Men are qualified for civil liberty in exact proportion of their disposition to put moral chains upon their own appetites... Society cannot exist unless a controlling power upon will and appetite be placed somewhere, and the less of it there is within, the more there must be without". The Hobbesian solution may be realistic, but it leaves open at least one important question: where do the shared rules that are enforced by the sovereign's meta-agency come from?

The origin of shared rules

5.1

The easiest explanation for the origin of this kind of shared rules, and the first one to be 'discovered', is that they were created in the mind of (and enacted by) some kind of real or virtual meta-agent. If this line of explanation is excluded, a second one (Hayek, 1973) may be contemplated: shared rules exist because, after having spontaneously emerged, they became functional to the society or the group. This approach, however, is problematic: it involves the explanation of a cause (shared rules as the cause of stable behavioural patterns) by its effects (the beneficial effects to the group or to the society) (Gilbert, 1995). It may easily be translated into the notion that all existing institutions are necessarily beneficial, and it leaves out the important case of rules of legislation that are deliberately enacted to achieve a certain purpose.

5.2

Alternative explanations for the origin of shared rules are built on the following set of premises: (a) Intelligent individuals may be able to recognise and evaluate some of the aggregated effects of shared rules; (b) These intelligent agents may even formulate theories that enable them to predict the outcome of alternative sets of shared rules and modes of meta-agency, and define preferences over these outcomes (Castelfranchi, 1998); (c) They may further engage in tacit or formal 'agreements' about institutional arrangements, that by influencing individual choices ensure the group's existence. From this perspective, the institutional framework is as much a spontaneous and non-intended result of a historic process as it is a product of a continuous negotiation among agents with possibly conflicting constitutional interests. Convergence of constitutional interests is a possible outcome of negotiation, but it seems more interesting to consider the existence of groups where a constitutional framework prevails without what might be defined as a voluntary agreement. As a matter of fact, it is not difficult to accept that one may submit to an order even if one's constitutional interests conflict with it - the prospect of living 'out in the wilderness' may simply seem unacceptable. The negotiation over the constitutional order may then be seen as a sort of game with temporary winners and losers, in which power counts. Morality can still be defined as a disposition to act in accordance with constitutional interests, but not as action in accordance with the prevailing shared rules (since the agent's constitutional interests may not converge). Institutional

change may now be explained not only as a result of constant adaptations in the individual minds, but as the outcome of changing balances of power. Rule enforcement, when domination enters the picture, can no longer be thought of as a neutral prerogative of some meta-agent. The Hobbesian story becomes a little less naïve.

Bounded rationality in a social context

6.1

In a formal way the environmental setting that we are interested in may be stated as follows: n agents live in a world where the system's state Y is determined by the set $\mathring{A} = (a_1, a_2,...,a_n)$ of actions of the individuals in some population. The function that maps $\mathring{A}$ into Y may be unknown to the agents and it may change in time, but given a state y of the system and the corresponding set $\mathring{A}$ of actions, every agent can assign credit to any action in $\mathring{A}$ using a function f_i that models the current state of his preferences and that maps $\mathring{A}$ into a set of evaluations $S'=(s'_1, s'_2,..., s'_n)$. The agents must recursively pick up an action from the set A of all feasible actions in a discrete sequence of time periods $t_1, t_2,...,t_T$.

6.2

The value of an action to any agent in this world is context dependent. It depends on the agent's preferences, on the function that maps $\mathring{A}$ into Y, and on the set $\mathring{A}$ of actions performed by all the other agents. This situation can be described as one of radical uncertainty: "Future events cannot be associated with probability distributions based on knowledge of the past" (Faber and Proops, 1998). This uncertainty arises from the behaviour of the other agents and from the aggregated behaviour of the system.

6.3

Simon's agents had limited perception and knowledge. Their choices were guided by evaluations of the expected consequences of actions, but they could neither perceive the whole range of admissible actions, nor perfectly compute the consequences of each of them. In placing Simon's agent in our social setting, a way of modelling limited perception is to assume that the actions observed in the present (the set $\mathring{A}_t$) are somehow *salient* to each individual. He will base his choice on the evaluation of actions belonging to this set. This neither implies that he must forget all the other actions that were observed in the past, nor that he is unable to perceive actions that were never observed, it is simply a consequence of imperfect knowledge. We are in fact assuming that the agent proceeds as if he believes that the distribution of actions in $\mathring{A}_{t+1}$ will be (at least) similar to the one observed in $\mathring{A}_t$, and that he credits the actions in $\mathring{A}_t$ under the assumption that y_{t+1} will be similar to y_t. There is obviously no rational justification for this (otherwise rationality wouldn't be limited), but if the system evolves in a smooth way the idea of a similar distribution of actions in consecutive time periods may not be totally arbitrary. It would be much harder to accept that the agent would also rely on evaluations of actions observed in the remote past, under totally different states of the system, or that he could evaluate actions that were never observed.

6.4

Let us have agent i in time period t deciding what to do in time $t+1$. The simplest way of modelling the decision procedure of one agent taking into account the preceding considerations may be the following one.

6.5

For the reasons given above we assume that the set $\mathring{A}_t$ is the one evaluated by the agents. Each individual has two alternatives when trying to reach a decision:

1. choose an action that 'looks promising' from the set of observed actions $\mathring{A}_t$, imitating it. In this case given the set S'_t the agent selects an action using a lottery where the probability of selection is somehow proportional to the credit assigned to each action in $\mathring{A}_t$. (Note that the agent might instead select the most credited action. However, this would be arbitrary since the credit assigned to each action is taken by the agent as mere indication. The lottery is then a device to model a choice under uncertainty in which 'intuition' guided by credit is in command.)
2. choose an action in A not included in $\mathring{A}_t$ in order to test it in t+1. In a world of limited knowledge and information there are reasons to be innovative. Opportunities may be hidden in a fog of uncertainty. This innovative move may be modelled in two ways: the agent may randomly modify the selected action in (1), or he may recombine the selected action with other actions selected by the same procedure.

6.6

When this decision procedure is simultaneously adopted by the n agents, the resulting process at the population level can be described as one of social learning: the population explores and adapts to an environment that includes all the agents' actions as a cause for its dynamics.

6.7

Behind such actions we are assuming particular rules as symbolic representations of those actions. Since in a given culture an agent is usually able to decode an action into underlying rules, the operations described above for the space of actions may also be viewed as operations in a space of rules, which may be stratified according to the different levels of choice. We also assume that the agents are informed of all the actions performed by other agents in a given time period, that they are able to assign credit to these actions (even though some of these were not directly experienced by them), and that they are able to code an observed action back into a 'program'.

6.8

Can we implement this tentative model of the agent's decision procedure? Can we simulate economic processes in the described setting with this type of agent?

The Genetic Algorithm as a model of socio-economic processes

7.1

When searching for an appropriate tool to model socio-economic processes, the Genetic Algorithm (GA), together with other evolutionary algorithms, appears to be a natural candidate (Holland, 1975; Goldberg, 1989). The appealing feature of the GA is that it may have behaviourally meaningful socio-economic interpretations. Arifovic (1991) mentioned two alternative interpretations, referred by Chattoe (1998) as a *mental interpretation* and as a *population interpretation*. They may be presented as follows: (a) In the *mental interpretation*, the population represents a single mind, i.e. each chromosome in the population represents a rule: "the frequency with which a given rule is represented in the population indicates the degree of credence attached to it" (Arifovic, 1991 :2) (b) In the *population interpretation* the population represents the active rule of each agent; the frequency of a given rule in the population indicates "the degree to which it is accepted in a population of agents" (Arifovic, 1991 :2).

7.2

Our interpretation of the GA differs from that of Arifovic on some specific points[2]. With modifications to the simple GA versions, it implements the model of bounded rational choice outlined above, combining elements of the mental and the population interpretations. The GA population represents a collection of sets of rules (even though each set of rules may correspond to an individual) associated with the set $\mathring{A}$ of actions; the *fitness function* is an individual credit

assigning function (not a system level function that determines the 'global' quality of the decision rules), and each agent is endowed with one that may be idiosyncratic; the *selection* operator implements the choice of actions from the set *Å* (imitation); the *mutation* operator corresponds to one type of innovative move; the *crossover* operator corresponds to the recombination of rules; depending on the innovative propensity of each agent the parameters, *probability of mutation* and *probability of crossover*, may vary. Given the model of the decision process outlined above, this modified GA may be summarised as follows[3]:

```
begin
  t<- 0
  randomly generate a population P(t)
  determine aggregate results for P(t)
  t<- t+1
  while not(stopping criteria) do
    begin
      for every agent j do
        begin
          assign credit to rule sets in P(t)
          select a rule set a_j1 from P(t)
          if r < prob_cross_j then
            begin
              select a rule set a_j2 from P(t)
              crossover(a_j1, a_j2) to a'_j
            end
          for every bit in a'_j
            if r <prob_mut_j then mutate that bit
        end
        determine aggregate results for P(t)
        t <- t+1
      end
end
```

Institutions as solutions to recurring problems in social interaction

8.1

Recall one of the questions put forward in the introduction: What is the role of institutions and how are they related to the aggregated performance of the socio-economic systems? It is useful to think of institutions as 'tools' or 'mechanisms' that provide solutions to recurring problems in social interaction (Vanberg, 1994). With the assistance of game theory and experimental economics, different types of recurring problems occurring in situations that involve no direct communication between agents may be identified. Their discussion and simulation with our model may help us to address the question posed.

Co-ordination (problem 1)

8.2

Let us first take a very simple co-ordination problem: "choose one of *n* colours; the payoff will grow with the number of players that will choose the colour you picked". If an experiment is carried out recurrently with the same players with choices and payoffs announced at the end of each repetition, we would most certainly observe that the choices would converge to one colour and stabilise there, even if the players were unable to communicate and agree on a common strategy. Starting from a

random choice of a colour, every player would understand that he should choose the alternative that, by chance, turned out to be the most often selected. Once choices converge, no single player would have any incentive to move to another alternative. The process of emergence of the equilibrium solution could be described in this case as a spontaneous one - the rule "choose green" or "choose red" would emerge and become a self-enforcing shared rule (convention) as if an invisible hand had driven the players to a social optimum.

8.3

This process may be simulated with the above specified model: each string stands for a choice-of-colour rule and each agent is endowed with the same function that assigns credit to the rules according to their absolute frequency in the population. The results of a typical run (100 players and 16 colours) show that (see figure 1), after an initial period where different colours still compete, the choices converge as expected to one, which turns out to be one of the most frequent random choices in the initial population. In nine out of ten runs of the simulation with different initial populations generated by different random generator seeds the procedure converged to one of the most frequent choices in the initial population and only once did it converge to the second most frequent. Convergence is not perfect due to mutation.[4]

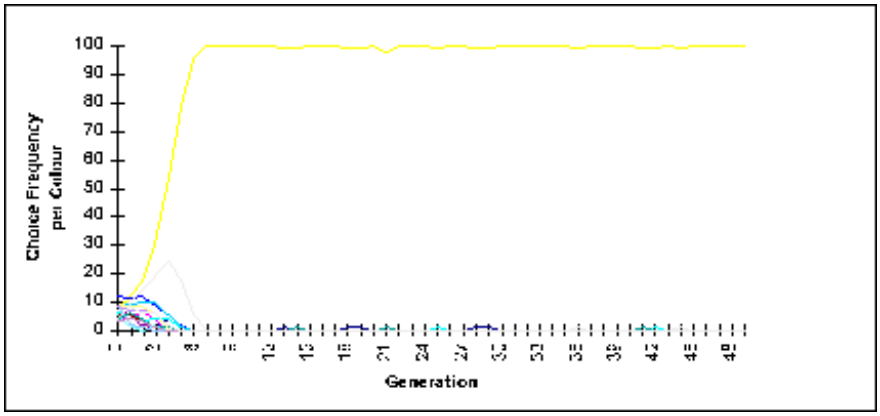

Figure 1: Co-ordination (problem 1): Choice frequency per colour through the simulation

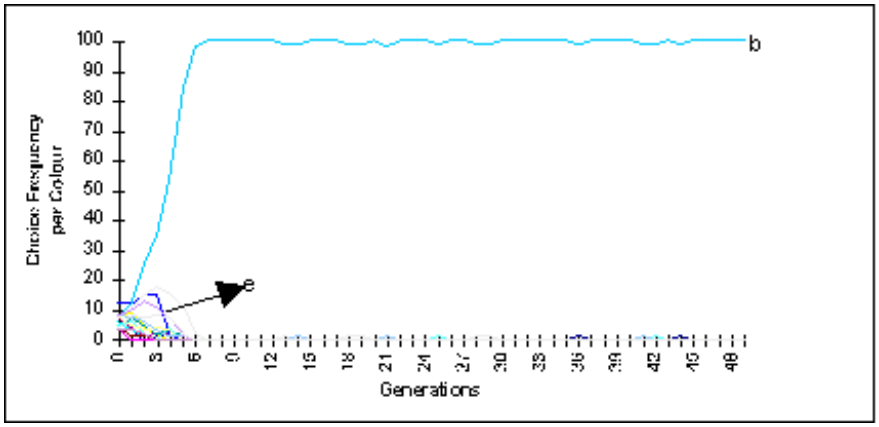

Figure 2: Co-ordination (problem 2): Choice frequency per colour through the simulation

Co-ordination (problem 2)

8.4

The players must choose among the same *n* colours. The payoffs will depend on the number of players choosing the same colour, but now it will also depend on the colour chosen. To co-ordinate on green (or on red) is better than any situation where choices are distributed, but to co-ordinate on red is better than to co-ordinate on green. In this case, as before, we could foresee that the players would co-ordinate. However it is much harder to be sure that they would always co-ordinate on the right colour. But once they co-ordinate on one colour (inferior or Pareto optimal), no individual

player will have an incentive to move to a different one. If they converge to an inferior colour they might only come to discover the best solution if they all moved simultaneously, and the chance that this might occur spontaneously is small, in particular if the number of players is large.

8.5

Simulating this process (100 players and 16 colours) with credit assigned by the function,

$$payoff(colour_i) = (Abs.\ Freq.\ colour_i) \times (1 + 0.2i)$$

we observe, as expected, a process of convergence. In the run reported in figure 2, choices that are Pareto inefficient (colours b and e) compete at the beginning of the simulation, and colour b that was not the most frequent in the initial population finally defeats colour e. Five of ten runs of the simulation with different initial populations converged to the Pareto optimal choice, the remaining simulations converged to a Pareto inferior outcome. Once again an invisible hand guides the agents, except that now it may lead them to an outcome that is not the best possible. We may conjecture, however, that if we would let the players discuss and agree on a joint strategy the chances of co-ordination in a Pareto optimal choice would increase.

8.6

In both cases, an institution - a convention - emerged as a non-intended result of the interaction of purpose-seeking agents in a typical 'invisible-hand' process. Both processes are instances of path dependency - a time irreversible outcome arises out of multiple possible causes including chance - and may be related to many processes that are observable in the real world. The first situation is often given as a possible account for the emergence of the 'keep to the right'- 'keep to the left' traffic regulation, the second was dwelt on by Davis (1985) in his famous story on QWERTY and more extensively by Brian Arthur (1994).

Co-operation

8.7

A third experimental situation may lead us further: a set of individuals, kept in isolation from each other, must post a contribution (from $0 to a pre-defined maximum) in an envelope, announcing the amount contained in it; the posted contributions are collected, summed by the experimenter and 'invested', giving rise to a collective payoff that must be apportioned among the individuals; the apportioning rule (known to the agents) stipulates that the share of the collective payoff must be proportional to the announced contributions (not to the posted contributions); the posted contributions and the corresponding announced contributions are subsequently made public (but not attributed to individuals); individual returns on investment are put by the experimenter into the corresponding envelopes and the envelopes are claimed by their owners.

8.8

This situation is related to the problem of team production (Alchian and Demsetz, 1972) but it also has common features with instances of public goods provision and collective action (Olson, 1965) and it can, in fact, be generalised to all situations where external effects are strong. For game theorists it is said to have a N-person repeated Prisoner's Dilemma (PD) structure.

8.9

Can we predict what is likely to happen after a number of repetitions of the experiment with the same experimental subjects? Ledyard (1995 :112) answers:

> There are many theories. One, the economic/game-theoretical prediction, is that no one will ever contribute anything. Each potential contributor will try to "free-ride" on the others. [...] Another theory, which I will call the sociological-psychological prediction, is that each subject will contribute something [...] it is some times claimed that

> altruism, social norms or group identification will lead each to contribute [...*x*...], the group optimal outcome. [...] Examination of the data reveals that neither theory is right.

As a matter of fact, the experimental evidence in similar cases shows that, with large groups, positive posted contributions are observable in the first rounds but free-riding soon emerges leading the group to levels of contribution that all agents consider undesirable.

8.10

The standard economic/game theoretical approach is unable to explain why positive contributions are observed in the first rounds of experiments: if I am self-interested, in the sense that I disregard the higher order obligation of contributing to collective goals and the prohibition of not telling the truth, and if I know that all the others disregard it in the same way, why should I contribute and be truthful, bearing the costs alone and having a benefit that is disproportional to my contribution? The functionalist sociological concept of *internalisation* might, in fact, explain the positive contribution in the first rounds. But would it explain the breakdown of the posted contributions observed with repetition? The question therefore is: What might be wrong with the 'economic/game-theoretical' and with the 'sociological-psychological' models?

8.11

In order to simulate this situation we have to take into account that in each period of time an agent must decide on his actual contribution and on his announced contribution. The shared rules: 'thou shall not lie' and 'thou shall contribute to the collective goal' are implicit. In order to take into account the preceding considerations on the 'hierarchy of feeling', the model of the agent must include not only a rule for the contribution that he will announce, but a second (special type) rule related to the degree to which he will comply with the shared rules. This rule set, attached to each agent, is implemented by coding one part of the 0/1 string (chromosome) as *announced contribution* and the other part as *moral disposition*; the *announced contribution* part of the string will decode into a real number between 0 and 50, and the *moral disposition* part to a real number between 0 and 1 [5].

8.12

The *posted contribution* of agent *i* is:

$$(A1)\ posted\ contribution_i = announced\ contribution_i \times moral\ disposition_i$$

The collective return on investment is given by:

$$(A2)\ collective\ return = 10 \times \sum_i posted\ contribution_i$$

The apportioning rule is:

$$(A3)\ return_i = \frac{announced\ contribution_i}{\sum_i announced\ contribution_i} \times collective\ return$$

The credit assignment function is:

$$(A4)\ credit_i = return_i - posted\ contribution_i$$

and the collective payoff is given by:

$$(A5)\ collective\ payoff = \sum_i return_i - \sum_i posted\ contribution_i$$

8.13

The results of a typical[6] simulation are shown in figure 3. Until the announced contributions reach their maximum level, the posted contributions increase as well, and after this they rapidly tend to zero, while the announced contributions are kept close to the maximum value. Due to the incentive

to free-ride, the initial moral disposition tends to erode with time. As a result, the collective payoff deteriorates, reaching the zero level around generation 120. After this, only occasional mutations (that might be interpreted as signalling intentions to contribute, conditional on the contribution of others) disturb the scenario of collective disaster.

8.14

The results are therefore consistent with the available experimental evidence: positive contributions are observed in the first rounds of the experiment but free riding tends to emerge leading the group to very low levels of contribution. A possible interpretation is that in spite of the initial moral disposition (which is unevenly distributed in the population), free-riding behaviour is rewarded and eventually invades the population of rules. No viable social order would be possible in this context, and the group would simply perish - such is the "tragedy of the commons". But let's not rush into easy conclusions: after all, in the real world this kind of collective action exists.

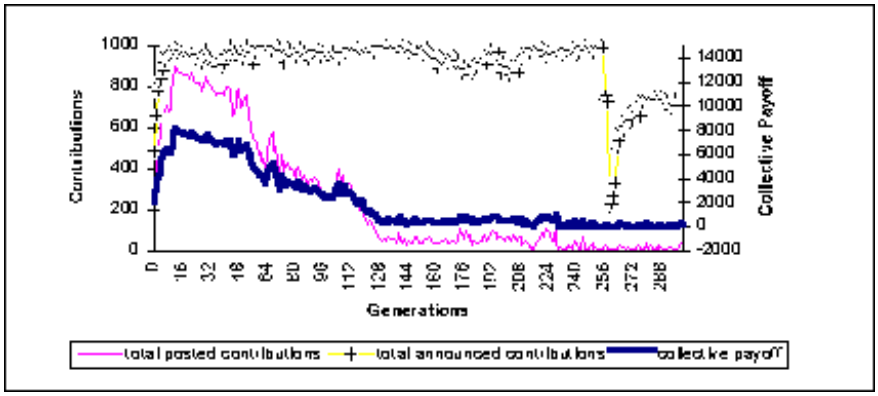

Figure 3: *probability of monitoring* = 0. Contributions and collective payoffs through the simulation

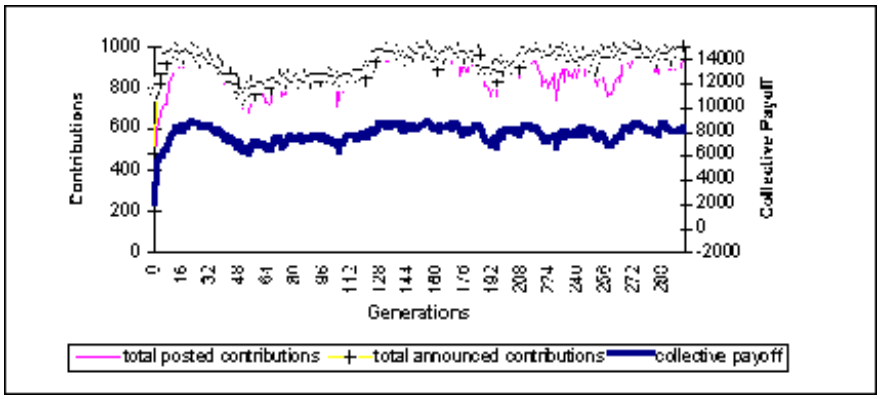

Figure 4: *probability of monitoring* = 1. Contributions and collective payoffs through the simulation

8.15

In this situation the Hobbesian solution would involve a bidding agreement to be enforced by the meta-agency of the sovereign. We will extend accordingly our simulation model, introducing a *meta-agent* with monitoring and sanctioning powers, while we continue to take as given the remaining institutional frame. Our only present aim is to test how the system would behave once the monitoring meta-agent is introduced by changing the initial setting of the experiment and the simulation: the experimenter (the meta-agent) may now decide to open some or (all) of the envelopes when they are handed to him; if an agent is found to have announced an amount that does not correspond to his contribution he will be sanctioned; his return on investment will now be determined by the following rule,

$$(A3')\ return_i = \frac{posted\ contribution_i}{\sum_i posted\ contribution_i} \times collective\ return \times moral\ disposition_i$$

with the implicit penalty reverting to the meta-agent and included in the collective payoff. The meta-agent chooses the individuals to be inspected by a simple rule: if *r* (a random real between 0 and 1) is lower than *probability of monitoring* (a parameter of the simulation), then agent *i*'s envelope will be opened.

8.16

The results of the simulation, with probability of monitoring set to 1 (all agents inspected) (see figure 4) show that by generation 20 the maximum value for the posted and announced contributions is reached and kept thereafter with some fluctuations. This means that the selective pressures exerted by the monitoring meta-agent successfully counteract free-riding and induce high levels of moral disposition. But does this mean that the Hobbesian meta-agency necessarily leads to Pareto optimal outcomes?

Extending the model

9.1

In the model that was previously presented, an apportioning rule was assumed: an unspecified deliberation process had led to the enactment of that rule. We are now interested in modelling: (a) the process through which the constitutional preferences of the agents may change as a result of their experience of the aggregated outcomes; (b) how the distribution of power within the group may be related with the constitutional design; and (c) how the constitutional regimes are related to the group's welfare.

9.2

The experimental setting and the model must once again be reformulated. Instead of announcing a contribution, the agent is now expected to abide by a minimum level of contribution (say 35). The meta-agent may decide to check if the envelope contains the minimum specified. If not, the agent will be sanctioned. In the model of the agent, a variable *size* is introduced representing the agent's *power*, which, in the context of this model, is related to the greater or lesser weight of each agent in the decision process that leads to the adoption of an apportioning rule, and (depending on the apportioning rule) it may influence the size of each agent's share of the collective benefits. The agent's model also includes a rule for *contribution* and *values* assigned to two alternative *apportioning rules* that allow the agent to choose among them. The apportioning rule is now chosen by a voting procedure in which an agent's *size* determines the weight of his vote. The *value* assigned by each agent to the constitutional rules is updated in every generation and is given by the agents' average individual pay-off under each rule's regime. The agent will vote on the rule that has greater value to him. The *size* of the agent is updated, assuming that in each generation a part of the individual payoff is 'capitalised'.

9.3

The collective return on investment is given by equation A2 and the apportioning rules are given by:

$$(B1)\ \text{Rule 1}\quad return_i = \frac{contribution_i}{\sum_i contribution} \times collective\ return$$

$$(B2)\ \text{Rule 2}\quad return_i = \frac{size_i}{\sum_i size_i} \times collective\ return$$

For non-monitored actions, under both regimes, the credit of action *i* to agent *j* is assigned by:

$$(B3)\quad credit_{i,j} = return_{i,j} - contribution_i$$

and for monitored actions with $contribution_i < 35$ we have

$$(B4)\ credit_{i,j} = return_{i,j} - contribution_i - penalty_i$$

with $penalty_i = 10 \times (35 - contribution_i)$.

The collective payoff is given by (A5), and the *size* of one agent is updated according to:

$$(B5)\ size_{t+1} = size_t + \frac{return_t - contribution_t - penalty_t}{100000}$$

The simulation includes a training period of 100 generations during which no voting takes place and which is used by the agents to explore the regimes of the two rules, assigning values to them. Rule 2 is experienced in the initial fifty generations and rule 1 throughout the next fifty. In generation 101, and every 20 generations after that, a vote takes place that may change the rule regime.

Simulation 1

9.4

All agents are created equal with size 10 and the probability of monitoring is set to 0.9. The results show (see figure 5) that in the first fifty generations (under rule 2) the total contributions and collective payoffs tend to decrease after the initial adjustment. When full monitoring is not possible, the apportioning of benefits in a way that is not proportional to contributions leads to an inefficient outcome. After generation fifty (under rule 1), the contributions and payoffs start to recover, reaching values that are close to the maximum amount. After generation 100, when voting starts, there is unanimity on rule 1 that is kept until the end of the run.

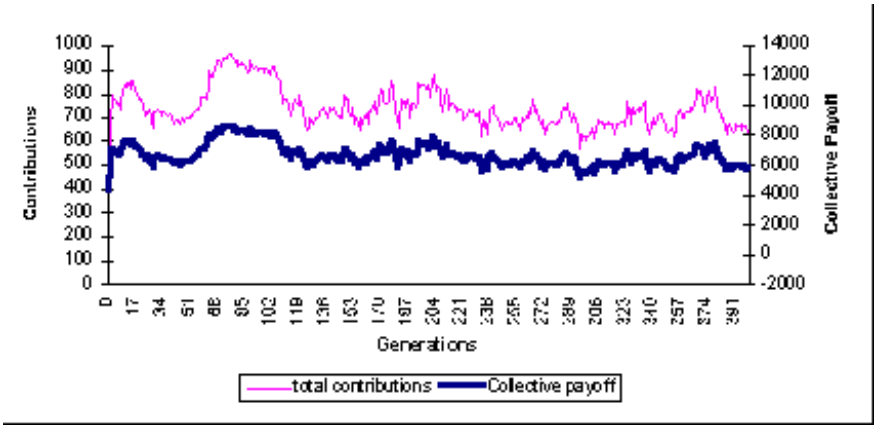

Figure 5: all agents created equal: contributions and collective payoffs through the simulation

Simulation 2

9.5

The size of each agent is now randomly generated varying between 0 and 20. The results (see figure 6) show that the comparatively bad performance of the first fifty generations (under rule 2) tends to improve under rule 1, between generations 50 and 100. However in generation 101, when it comes to voting, rule 2 wins (rule 2 has a majority of votes even though it does not have a majority of voters; see figure 7). In point of fact, rule 2 performs well for large agents and badly for small ones - the correlation between the *value* of rule 2 and the *size* of the agent is almost perfect. After generation 100 (under rule 2) the overall pattern of the collective payoffs and contributions is inefficient and rather unstable.

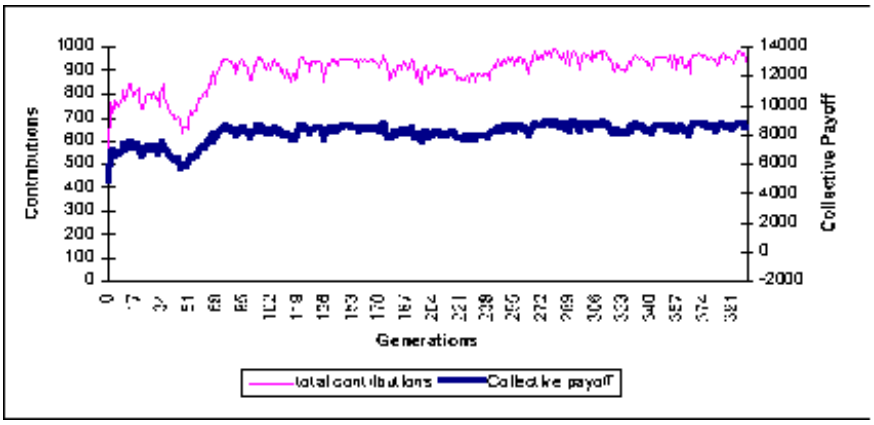

Figure 6: distributed power: contributions and collective payoffs through the simulation

9.6

These results suggest that a rule regime that is not beneficial to the group may persist given an unbalanced power distribution and high levels of monitoring.

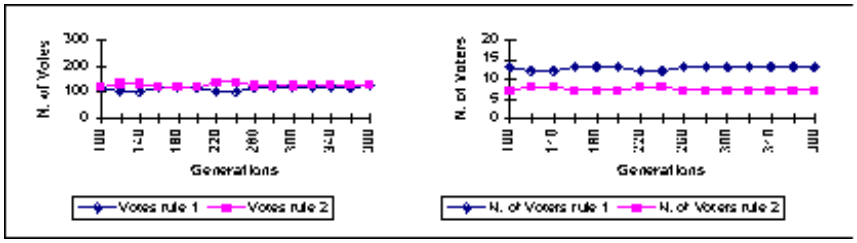

Figure 7: distributed power: number of votes and voters

Conclusion

10.1

We started with two initial questions: What is the role of institutions and how are they related to the aggregated performance of the socio-economic systems? What is the relation between institutions and the behaviour of the agents? Viewing institutions as 'tools' or 'mechanisms' that provide solutions to recurring problems in social interaction, we explored different environments that have in common the anonymous nature of the interaction and the absence of direct communication between agents.

10.2

Concerning the first question, we conjectured that under some circumstances conventions might spontaneously come into being that provide the basis for social life. These conventions would be self-enforcing; that is, they would be reproduced by the practices of the agents without the interference of any 'external' meta-agency. However, we were led to the conclusion that even in this case no support could be given to the deep rooted belief in a 'naturally' beneficial outcome generated by an invisible hand process - the emergent convention could correspond to a Pareto inferior outcome.

10.3

We moved then to contexts where spontaneous processes seem to generate outcomes that are not compatible with the existence of the group. After confronting this possibility with optimistic prospects of an anarchic order, we considered the more familiar reality of a social order based on a contract and an enforcing meta-agency subject to the collective choices of the group. However, the visible hand of meta-agency, and the underlying social choice, also seem to be no guarantee of efficiency: the mode of meta-agency and the constitutional rules that it enforces affect in a critical way the performance of the group. Worse, the evaluation of the constitutional regimes by the agents might differ, depending not only on the aggregate results, but on the particular situation of an agent

within society. The results of simulation with this model suggested that the social choice over alternative constitutional rules, when biased by the distribution of power within the group, may lead to rule regimes that, although generating inferior outcomes for the group, may (or may not) be sustained in time, through coercion.

10.4

Addressing the second question, we started from a model of a bounded rational agent living in society. That model was extended to consider situations where an agent must balance his situational action interest with his moral obligations. We were forced to conclude that when the incentive structure favours free-riding the moral disposition may not be sustainable. However, intelligent agents may be able to recognise the links between their personal fates and group welfare, and settle on 'agreements' which, for lack "of moral chains within", may be enforced by "a controlling power, without". For this, a new extension of the model of the bounded rational agent was necessary in order to include preferences and choices over alternative shared rules, and a social choice procedure.

10.5

All through this research, simulation was used as a device to check the consistency of the models and to derive their implications. It imposed a discipline on the modelling activity and very often provided insights and gave rise to new ideas to be explored (in particular when the human simulator was confronted with unexpected results). The implemented simulation model served its purposes well but it remains open for new extensions, for instance, situations of interacting multiple populations, and agents with different preference structures. In the future we intend to explore the extent to which some of the assumptions of the model may be relaxed by implementations of the mental interpretation of the GA and by other Evolutionary Algorithms, in particular when face-to-face interaction involving communication is considered.

Acknowledgements

This research was partially funded by the Fundação para a Ciência e Tecnologia under the PRAXIS XXI Programme (Project SARA 2/2.1/TIT/1662/95) and supported by DINAMIA/ISCTE (Centro de Estudos Sobre a Mudança Socioeconómica). The authors would like to thank the anonymous referees for their generous and valuable criticism and suggestions.

Notes

[1] Three main trends can be identified:

1. one with roots in neoclassic economics, labelled as 'New Institutional Economics' (Williamson, 1985; North, 1990; Olson 1965),
2. one founded in the Austrian tradition and represented by the more recent research of Hayek (for instance, 1973), and
3. one inspired by the 'old' American institutionalism of Veblen and Commons (Hodgson, 1988, 1993).

[2] Arifovic's use of the GA has been discussed by Chattoe (1998) and in Caldas and Coelho (1999).

[3] P(t) stands for the population in generation t, r is a uniformly distributed random number between 0 and 1; $prob\_cross_j$ and $prob\_mut_j$ are the parameters that set the probability of crossover of a selected chromosome and the probability of mutation of each single bit for agent j.

[4] Even though in this particular context mutation is hardly justifiable in rational terms, note that instability in convergence is also observable with human beings in similar experimental contexts; boredom or insufficient understanding of the game situation are usually the explanations given by researchers. The authors thank an anonymous referee for pointing out that an important feature, that has been experimentally observed is not captured by the model: some strategies, for instance 'black' and 'white' strategies, may be prominent. In multiple runs in the laboratory they would be observed as outcomes with a non-random frequency.

[5] In all the simulations that follow, the *Population Size* is 20, the *Probability of Crossover* is 0.5 and the *Probability of Mutation* is 0.01, for all agents.

[6] Different initial populations were generated using various random generator seeds. The observed overall pattern of outcome is common to all.

References

AKERLOF, George A. (1980), "A Theory of Social Custom, of Which Unemployment May Be One Consequence". *The Quarterly Journal of Economics*, Vol. XCIV, n. 4, 749-75.

ALCHIAN, Armen A. and Demsets H. (1972), "Production, Information Costs, and Economic Organization". *American Economic Review*, vol. 62, December 1972, pp 777-795.

ARIFOVIC, Jasmina (1991), *Learning by Genetic Algorithms in Economic Environments*. Doctoral Dissertation, Department of Economics, Chicago: University of Chicago.

ARTHUR, Brian (1994), *Increasing Returns and Path Dependence in the Economy*. Ann Arbor, Michigan: University of Michigan Press.

CALDAS, José C. and Coelho H. (1994), "The Simulation of Trade in Oligopolistic Markets". In Doran J. and Gilbert, N. (Eds.) *Simulating Societies: The Computer Simulation of Social Phenomena*. London: UCL Press.

CALDAS, José C. and Coelho H. (1999), "Agents, Groups and Institutions", Working Paper, DINAMIA/ISCTE, (forthcoming).

CASTELFRANCHI, Cristiano (1998). "Simulating with Cognitive Agents: The Importance of Cognitive Emergence". Pre-Proceedings MABS'98 (Multi-agent systems and Agent-Based Simulation). July 4-6, 1998, Cité des Sciences - La Villette, Paris, France.

CHATTOE, Edmund (1998), "Just How (Un)realistic are Evolutionary Algorithms as Representations of Social Processes". *Journal of Artificial Societies and Social Simulation (JASSS)*, vol. 1, no. 3, http://jasss.soc.surrey.ac.uk/1/3/2.html.

CONTE, Rosario and Castelfranchi, C., (1995), *Cognitive and Social Action*. London: UCL Press.

DAVIS, Paul (1985), "Clio and the Economics of QWERTY". *American Economic Review*, n. 75, pp. 332-337.

FABER, Malte and Proops, John, *Evolution, Time, Production and the Environment*. Third Revised and Enlarged Ed. Berlin: Spriger. EDGEWORTH, F.Y. (1881), *Mathematical Psychics*. London: Kegan Paul.

GILBERT, Nigel (1995), "Simulation: an emergent perspective". 7/28/98, http://www.soc.surrey.ac.uk/research/simsoc/tutorial.html.

GOLDBERG, David E. (1989), *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading, Massachusets: Addison-Wesley.

HAYEK, Friedrich A. (1973), *Law, Legislation and Liberty*, Vol. 1 - *Rules and Order*. Chicago: The University of Chicago Press.

HEAP, S.H., Hollis, M., Lyons, B., Sugden, R., and Weale A. (1992), *The Theory of Choice: A Critical Guide*. Oxford UK: Blackwell.

HODGSON, Geoffrey M. (1988), *Economics and Institutions: A Manifesto for a Modern Institutional Economics*. Cambridge, UK: Polity Press.

HODGSON, Geoffrey M. (1993), *Economics and Evolution: Bringing Life Back Into Economics*. Cambridge, UK: Polity Press.

HOLLAND, John H. (1975), *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. Second edition, Cambridge, Massachusetts: The MIT Press, 1992.

JEVONS, W. S. (1871), *The Theory of Political Economy*. Reprints of Economic Classics, New York: Augustus M. Kelley, 1965.

LEDYARD, John O. (1995), "Public Goods: A Survey of Experimental Research". In John Kagel and Alvin Roth (Eds.), *The Handbook of Experimental Economics*, Priceton, New Jersey: Princeton University Press.

MACY, Michael W. (1998), "Social Order in Artificial Worlds". *Journal of Artificial Societies and Social Simulation (JASSS)*, vol. 1, no. 1, http://jasss.soc.surrey.ac.uk/1/3/2.html.

MARGOLIS, Howard (1991), "A New Model of Rational Choice". In *Rational Choice Theory*, Aldershot: Edward Elgar Publishing.

NORTH, Douglass C. (1990), *Institutions, Institutional Change and Economic Performance*. 5th edition, Cambridge: Cambridge University Press.

OLSON, Mancur (1965), *The Logic of Collective Action*. Cambridge, Massachsets: Harvard University Press.

SIMON, Herbert (1955), "A Behavioral Model of Rational Choice". *Quarterly Journal of Economics*, n. 69, pp. 99-118.

SIMON, Herbert (1956), "Rational Choice and the Structure of the Environment". *Psychological Review*, n. 63, pp. 129-138.

SUGDEN, Robert (1986), *The Economics of Rights, Co-operation, and Welfare*. Oxford: Blackwell.

VANBERG, Viktor J. (1994), *Rules and Choice in Economics*. London: Routledge.

WILLIAMSON, Oliver E. (1985), *The Economic Institutions of Capitalism*. First Paperback Edition, New York: The Free Press, 1987.

Return to Contents of this issue