INDIVIDUAL ALIENATION AND SYSTEMS INTELLIGENCE

H. Atlan*, J.-P. Dupuy and M. Koppel*****

**Department of Medical Biophysics, Hadassah University Hospital, Jerusalem, Israel*
***CREA, Ecole Polytechnique, 1 rue Descartes, 75005 Paris, France*
****Department of Mathematics, Bar-Ilan University, Ramat-Gan, Israel*

Abstract. H. von Foerster's conjecture: the more trivially connected the elements of a system, the less influence they will have on the system (= the more "alienated" they will be).
We show that this conjecture provides a firmer foundation for the concept of equilibrium in economics and in game theory; and in particular that it justifies the hypothesis of the alienation of individual agents that is inherent in any concept of equilibrium. We apply the techniques of information theory to probabilistic automata in order to formalize and to prove this conjecture.

Keywords. General economic equilibrium; intelligent systems modeling; probabilistic automata; information theory; complexity; alienation.

THE LOGICAL AND EPISTEMOLOGICAL FOUNDATIONS OF THE CONCEPT OF EQUILIBRIUM IN ECONOMICS AND IN THE SOCIAL SCIENCES

With the concept of general equilibrium, economists believe they have resolved in theory the problem which is at the heart of every modern analysis of society: how to conceive of society as a self-sufficient entity, informationally and operationally closed, owing to no external source the principle of its functioning and achievements; in terms borrowed from systems theory: how to conceive of society as a self-organizing system. The economic solution to this problem also satisfies in principle the requirements of methodological individualism: society being nothing other than a composition effect produced by interactions among individuals, there exists no locus which would constitute the center of social regulation; the regulation mechanism is "distributed" over the entire set of individuals--nowhere is it localizable.

Economists sometimes turn to computer metaphors in order to describe this conception of the social order. The prices which permit the decentralized achievement of an equilibrium (and of an optimum) state of the economy are calculated by the market itself, which in this way functions as a huge "macro-computer" that hasn't been manufactured nor even less programmed by anyone: a self-made computer and a self-programming program, in some sense. Today, in artificial intelligence, so-called neo-connectionism has similar objectives, and one can therefore say that general economic equilibrium is already the prototype of a neo-connectionist conception of society.

However, the most lucid economists recognize that at the present time this research program still designates a goal to be reached more than a definitive attainment. The principal obstacle to be overcome remains the famous figure of the walrasian auctioneer, that stubborn residue of an exteriority which one would well like to do totally without. The most vexing part of the business is that, as Schotter (1983) writes, "the problem with this assumption is not its lack of realism, because, although these fictitious auctioneers clearly do not exist, it can be said that markets do function as if they really did." Indeed, the auctioneer is here but the symbolic personification of a hypothesis that is crucial to the overall coherence of the model: "all agents behave as price takers and maximize the value of their objective function, taking these prices as given by a deus ex machina known as the fictitious auctioneer." The true exteriority that subsists is thus that of the prices in relation to the agents. It is all the more paradoxical in that the designer of the model sees quite clearly from his own external vantage point that it is the agents who, collectively, determine the prices. This situation causes some Marxists to say that the neoclassical model of the market suffers from a dyed-in-the-wool internal contradiction, since it supposes that "the firm does not have the ability to modify the market prices but can only adapt itself to them; now this contradicts the general hypothesis that each economic agent contributes through his supply and his demand to the formation of prices." (Godelier, 1968)

The justifications advanced by economists to explain this exteriority of prices are varied and well-known. By general admission, none is really convincing and decisive. The simplest and most brutal consists in postulating that it is a matter of a hypothesis about the agents' representations which, as pure hypothesis, need not be discussed any further, which may or may not be satisfied depending on the context, and which, when it is, defines a situation called "perfect competition". The problem with this first justification is that it is open to the Marxist criticism: the market could only function, then, at the cost of an

"alienation" of the agents--they would not see what the theorist alone sees, namely that the obstacle they run up against (the prices) is one that they themselves have raised.

A second and more common justification consists in asserting that the agents are right to suppose they have no influence on the prices, for such is indeed the case: each is too small, in relation to the market as a whole, for his action to have any observable effect. The difficulty here is to justify why the agents satisfy themselves with this situation of atomization and don't decide to regroup their forces by forming coalitions. This brings us to the third justification, the most sophisticated one. In the case of agents who are (infinitely) numerous, if they imagine all the ways of forming coalitions and reject each time the states of the economy that would damage the coalition under consideration, the only states remaining would be those that coincide with competitive equilibria. This famous result of the work of Debreu and Scarf (1963) on the asymptotic equivalence between the concept of core and that of competitive equilibrium would thus establish the equivalence between an irrational, "alienated" and short-sighted mode of behavior (considering prices as given) and the most rational and well-informed one (placing oneself at the core). But two substantial difficulties arise here. First, the information cost necessary in coalition-forming very quickly becomes prohibitive as the number of agents increases. Next, one can show that once it is known to the agents, the concept of core becomes a self-defeating concept (Morgenstern and Schwödiauer, 1976): certain traders in the economy have an interest in stopping the recontracting process at some social state outside of the core by forming a cartel. Consequently, one can say that perfectly rational agents, well-versed in all of these theoretical results, who wish to coordinate their actions in a competitive equilibrium, will be much better off feigning belief in what they know quite well to be untrue, namely that the prices are given. But is this self-delusion compatible with our habitual notion of rationality?

These difficulties clearly do not stem from the specific nature of the equilibrium variables, in this case prices. They are inherent in the very concept of equilibrium when it is imported from its birthplace in mechanics and physics into the social sciences. This has been made evident by, among other things, researches on the microeconomic foundations of macroeconomics (disequilibrium theories) and, above all, on rational expectations. The equilibrium variables here can be of any nature at all. The principal lesson, moreover, of these diverse models was this (even if the "realist" position still has its advocates): the representations of agents in an equilibrium are neither true nor false, they are self-fulfilling by the intermediary of the actions they spawn (Dupuy, 1982; Guesnerie, 1983; Orléan, 1986). In the language of the theory of autonomous, autopoietic or self-organizing systems (Atlan, 1979; Varela, 1979) as it has developed in theoretical biology, one would say: an equilibrium is a fixed point (or "eigen-behavior") of the operator that describes the organizational and informational closure of the system. The problem in human affairs is that, before explaining how the circle closes in on itself, one must justify how and where it opened--which takes us back to the question of the exteriority of equilibrium variables for the agents.

The concept of equilibrium in game theory is no more securely grounded, but this theory provides us with two valuable insights. In contrast to the Walrasian model of the market, it sets on the stage agents for whom the exercise of rationality implies that they put themselves through their imagination in the place of others. When they do so, they see that the others are doing the same in regard to them, and the resulting game of mirrors within mirrors is in principle without limit. What blocks this infinite regress is, precisely, the concept of equilibrium--the fact that beyond a certain level, each agent takes his own supposition concerning the other as a given, and not as a relation itself susceptible to being reflected back again in the mirror of the other. "Alienation" has therefore a fuctional role here, it arbitrarily puts an end to the potentially boundless game of mutual fascination and unbridled suspicion. In addition, the interpretation of equilibrium in terms of self-fulfilling representations allows us in this case to say that what is self-fulfilling and therefore "self-founding" in an equilibrium is not only the expected values of the variables, but also their nature; it is not only then some one particular equilibrium, but also its type, characterized by the level at which the expectations stop (Nash, Stackelberg, etc.) (Walliser, 1985).

This last observation suggests a path for research on the problem that concerns us: how to provide a foundation for the hypothesis of individual alienation inherent in any concept of equilibrium. The idea is to suppose that what is self-fulfilling and "self-founding" in an equilibrium is not only the value of the variables and the type of equilibrium, but also the hypothesis of exteriority that the agents make regarding these variables. If we can show this, we will then have succeeded, not in eliminating exteriority, but in endogenizing it--by accounting for it in terms of circular causality, after the fashion of self-fulfilling prophecies. Considerations of a more general nature suggest that it is impossible to conceive of the autonomy of society without calling on this idea of a pseudo-exteriority or endogenized exteriority (cf. the notions of endogenous fixed point and of self-transcendance in Dupuy, 1986).

VON FOERSTER'S CONJECTURE

Introduction

We have availed ourselves of a conjecture of von Foerster's, formulated in 1976 in the context of systems theory and automata networks (Dupuy, Robert, 1976; Dupuy, 1982). It applies to the class of systems in which the actions of a set of individual agents determine the very state of the system which in turn serves as the reference point for these same actions. This circular causality between agents and environment is evidently at the heart of the concept of economic equilibrium that has figured in our discussion, but it is equally common to a number of interesting social situations: crowd and panic phenomena (Dupuy, 1983), the choice of transportation and itinerary by an urban dweller, diploma-based competition and the devaluation of degrees (Boudon, 1973), etc.

The conjecture is that the more the elements of a system are "trivially" connected, the less will be their influence on its overall behavior; therefore, the more will they observe that the environment is untouched by their actions, as if external to them; in other words, the more will they be alienated. By "trivially" connected, von Foerster means that the influence of the state of the system on the action of the elements takes the form of a rigid, univocal determination.

The fact that individual behaviors must be "complex" (in the sense of "non-trivial", un-rigid) for the agents to have a chance to exert an

influence on the system may appear paradoxical insofar as the overall behavior of the system is all the more predictable for an external observer when these behaviors are less complex. Here one recognizes the crucial importance of the observer's position.

Before showing how we have formalized and proved this conjecture, let's look at how it will resolve our problem. Suppose the agents conceive of certain environmental variables as being external to them and indifferent to their actions. As they are by hypothesis maximizing agents, they are going to conduct themselves in trivial fashion (for example, with the prices known, their behavior is determined). Note that if they were not alienated, despite their being maximizers they would lose themselves in the endlessly-refracted mirror images that define their relations with others, and their behavior would be highly indeterminate. But here, being alienated, they behave trivially. If the conjecture is correct, they are going to verify that their starting hypothesis was well-founded, namely that they have no influence on the environment. Alienation is therefore a form of self-fulfilling representation. Note the importance of the hypothesis of maximizing behavior in the obtainment of this result. Between the concept of equilibrium and that of individual rationality there thus exists a bond of coherence that remained undisclosed until now.

Formalization and Theorem (Koppel, Atlan and Dupuy, 1986)

The theoretical framework is the theory of probabilistic automata. We draw on concepts from Shannon's information theory, with dynamic sources.

Suppose a probabilistic cellular automaton S. The state of S at the moment t determines, for each cell in S, the probability of its being in one or another of its possible states at the moment t + 1. The "environment", D, is defined as the largest subset of S such that, for all t, the state of S at t determines the state of each cell in D at t + 1. The cells in S which are not in D, written $A_1, \ldots, A_n$, are "free agents" in S (thus, D is a deterministic automaton, of which the inputs are the states of the free agents; $\{A_1, \ldots, A_n\}$ is a probabilistic automaton having D for input).

We make use of the following definitions and notations.

The mutual information of two sources is defined as

$$I(X_1 : X_2) = H(X_1) - H(X_1 | X_2), \tag{1}$$

and the mutual information of two sources given a third source is defined as

$$I(X_1 : X_2 | X_3) = H(X_1 | X_3) - H(X_1 | X_2, X_3). \tag{2}$$

This then gives us the fundamental equation

$$I(X_1 : X_2 | X_3) = I(X_2 : X_1 | X_3) = H(X_1 | X_3) + H(X_2 | X_3) - H(X_1, X_2 | X_3) \tag{3}$$

The total mutual information of three sources is defined as

$$I(\{X_1, X_2, X_3\}) = H(X_1) + H(X_2) + H(X_3) - H(X_1, X_2, X_3). \tag{4}$$

The intersecting mutual information of three sources is defined as

$$I(X_1 : X_2 : X_3) = I(X_1 : X_2) - I(X_1 : X_2 | X_3) \tag{5}$$

This then gives us

$$I(X_1 : X_2 : X_3) = I(X_1 : X_3 : X_2), \tag{6}$$

and equations in the same manner for all the permutations.

We are now able to define the mutual information between the mutual information of a set of sources, on one hand, and a further source on the other hand:

$$I(\{A_1, A_2, A_3\} : A_4) = I(\{A_1, A_2, A_3\}) - I(\{A_1, A_2, A_3\} | A_4). \tag{7}$$

Going back to our automaton S, we let S^t, D^t and A^t be the "sources" S, D and A, respectively, at the moment t. For the sake of convenience, we let $F = \{A_2, \ldots, A_n\}$, and let $A_1^{i,t}$, $F^{i,t}$ and $S^{i,t}$ designate the sets $\{A_1^i, \ldots, A_1^t\}$, $\{F^i, \ldots, F^t\}$ and $\{S^i, \ldots, S^t\}$, respectively.

Our object is to formalize the influence of the free agent 1, thought of as isolated from the others, on the environment D. More precisely, we establish a value for the influence of $A_1^{1,t}$ on D^{t+1}, which we write as $C(A_1^{1,t} \rightarrow D^{t+1})$, as follows:

$$C(A_1^{1,t} \rightarrow D^{t+1}) = I(A_1^{1,t} : D^{t+1} | F^{1,t}) + \sum_{i \leq t-1} I(A_1^i : F^{i+1,t} : D^{t+1} | S^{i-1}). \tag{8}$$

The first term, $I(A_1^{1,t} : D^{t+1} | F^{1,t})$, represents the information on D^{t+1} that is contained in $A_1^{1,t}$ but not in $F^{1,t}$. The second term represents the information on D^{t+1} that is contained in $A_1^{1,t}$ and in $F^{1,t}$, but such that it is first furnished by A_1: F only contains this information by virtue of having "copied" it from A_1. In other words, the total influence on the environment is the sum of a direct influence and of an indirect influence relayed by the influence on the other agents.

As for the complexity (non-triviality) of A_1 at the moment t, it appears quite simply as $H(A_1^t / S^{t-1})$.

We then obtain the following theorem, the demonstration of which proves von Foerster's conjecture as we have formulated it.

Theorem: $$C(A_1^{1,t} \rightarrow D^{t+1}) = \sum_{i=1}^{t} H(A_1^i | S^{i-1}) - \sum_{i=1}^{t} H(A_1^i | S^{i-1}, D^{t+1}). \tag{9}$$

The complexity of A_1 thus constitutes the upper bound on the influence of $A_1^{1,t}$ on D^{t+1} (this bound is indeed the smallest since it can be attained for an appropriate environment when the second term equals zero). For each free agent, then, weak complexity (or, what is the same thing, strong triviality) implies a weak influence on the environment (or: strong alienation). Q.E.D.

A Remark on the Concept of Influence

Consider the following example. At any time t let the free agent A_1 be in any of the states (a, b, c) and the environment D be in one of the states

(x, y). Other free agents in the state description are immaterial. The environment evolves as follows:

$$D^{t+1} = \begin{cases} x, \text{ if } A_1^t = a \text{ or } b \\ y, \text{ if } A_1^t = c \end{cases} \quad (10)$$

(Thus D^{t+1} is completely determined by S^t, in particular by A_1^t.)

Also:

$$P(A_1^{t+1} = a | S^t) = P(A_1^{t+1} = b | S^t) = (1 - \varepsilon)/2, \quad (11)$$

and

$$P(A_1^{t+1} = c | S^t) = \varepsilon. \quad (12)$$

Then, taking t = 1, clearly (logs to base 2)

$$H(A_1^1 | S^0) = -\varepsilon \log \varepsilon - (1 - \varepsilon)\log(1 - \varepsilon) + (1 - \varepsilon). \quad (13)$$

Because D^{t+1} determines whether A_1^t is in (a, b) or (c):

$$P\{(a, b) | D = x\} = 1; \quad (14)$$

and a, b are equiprobable:

$$H(A_1^1 | S^0, D^2) = -(1 - \varepsilon)\log(1/2) = 1 - \varepsilon. \quad (15)$$

Thus by the theorem the "influence" of A_1^1 on D^2 is

$$C(A_1^1 \rightarrow D^2) = H(A_1^1 | S^0) - H(A_1^1 | S^0, D^2) = -\varepsilon \log \varepsilon - (1 - \varepsilon)\log(1 - \varepsilon). \quad (16)$$

It can be made arbitrarily small by making ε small.

This result may seem paradoxical since, by hypothesis, the state of A_1^1 completely determines the state of D^2 irrespective of ε. But this influence of A_1^1 on D^2, which is in fact a determinism, is but a potential influence. The influence C under consideration is not strictly speaking a perceived or subjective influence, it is perfectly objective; but it depends crucially on the history of actions accomplished by the agent (and, in the general case, by the others). This simple example makes clear how the agent's relative triviality, by diverting him from taking certain actions (here, c), keeps him from exploring fully his potential influence. Although he has here in his power total control of the environment, he finds himself in a situation where, whether he does a or b, the environment stays fixed at x: his effective influence is nil.

CONCLUSION

This model does not claim to be a substitute for the diverse variants of the economic model of general equilibrium. It is situated on another level: it doesn't seek to describe the functioning of a reality, but that of a concept--the concept of equilibrium.

Nor does it claim to decide among the myriad numbers and natures of the equilibria that economists discover in exploring the theoretic possibilities of their models. It is on the contrary a corollary of the foregoing considerations that the richness of this multiplicity should be credited to the concept of equilibrium itself. In human affairs, the circles formed by interpersonal relations close in on themselves in many possible ways, the arbitrariness of which is largely irreducible. The "trivialization" inherent in all life in society appears as a facilitating condition.

It seems finally that a certain dose of opacity, of misapprehension and of reification is a necessary condition for the emergence of any social equilibrium.

Acknowledgements

This research was possible thanks to funding from the Integrated Research Action "Sciences de la Communication" of the C.N.R.S.

REFERENCES

Atlan, H. (1979). Entre le Cristal et la Fumée. Le Seuil, Paris.

Boudon, R. (1973). L'Inégalité des Chances. Colin, Paris.

Debreu, G., and H. Scarf (1963). A limit theorem of the core of an economy. International Economic Review, 4, 234-246.

Dupuy, J.-P. (1982). Ordres et Désordres. Le Seuil, Paris.

Dupuy, J.-P. (1983). De l'économie considérée comme théorie de la foule. Stanford French Review, VII, 245-263.

Dupuy, J.-P. (1986). L'Autonomie du social. In Encyclopédie Philosophique. P.U.F., Paris. Forthcoming.

Dupuy, J.-P., and J. Robert (1976). La Trahison de l'Opulence. P.U.F., Paris.

Godelier, M. (1968). Rationalité et Irrationalité en Economie. Maspero, Paris.

Guesnerie, R. (1983). L'influence des représentations des acteurs sur les faits économiques et sociaux objectivement constatables: une contribution introductive. In P. Dumouchel and J.-P. Dupuy (Eds.), L'Auto-Organisation. Le Seuil, Paris.

Koppel, M., H. Atlan, and J.-P. Dupuy (1986). Von Foerster's conjecture--trivial machines and alienation in systems. Forthcoming.

Morgenstern, O., and G. Schwödiauer (1976). Competition and collusion in bilateral markets. Zeitschrift für Nationalökonomie, 36, 217-245.

Orléan, A. (1986). Mimétisme et anticipations rationnelles: une perspective keynsienne. Recherches Economiques de Louvain, 1. Forthcoming.

Schotter, A. (1983). Why take a game theoretical approach to economics? Institutions, economics and game theory. Economie Appliquée, 36, 673-695.

Varela, F. (1979). Principles of Biological Autonomy. Elsevier North Holland, New York.

Walliser, B. (1985). Anticipations, Equilibres et Rationalité Economique. Calmann-Lévy, Paris.