Rationality in Multi-Agent Systems

KEN BINMORE[1], CRISTIANO CASTELFRANCHI[2], JAMES DORAN[3], and MICHAEL WOOLDRIDGE[4]

[1]*ELSE, Department of Economics, University College London, Gower Street, London WC1E 6BT, UK*
(email: K.Binmore@ucl.ac.uk)
[2]*IP-CNR, Italian National Research Council, Viale Marx 15, I-00137 Rome, Italy*
(email: cris@pscs2.irmkant.rm.cnr.it)
[3]*Department of Computer Science, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK*
(email: doraj@sx.ac.uk)
[4]*Department of Electronic Engineering, Queen Mary & Westfield College, University of London, London E1 4NS, UK*
(email: M.J.Wooldridge@elec.qmw.ac.uk)

Abstract

This report is the result of a panel discussion at the *Second UK Workshop on Foundations of Multi-Agent Systems (FOMAS-97)*. All members of the panel are authors, listed alphabetically.

1 Introduction

Since the publication of Herbert Simon's *Sciences of the Artificial* (Simon, 1981), the concept of *rationality* and *rational action* has been central to the study of Artificial Intelligence (AI) and, more recently, to the study of intelligent agents and multi-agent systems. One (of the many possible) ways of defining agents is as *rational decision makers*, and indeed, one recent textbook on AI defines AI itself as the enterprise of constructing such rational agents (Russell and Norvig, 1995). The concept of rationality is itself quite a simple one, and yet the theories that emerge from the concept of rational action turn out to be quite profound. Put very crudely, we might say an agent is being rational if it tends to act in its own best interest.

In economics and game theory (Binmore, 1992), this concept is formalized by attributing to every actor a *utility function*, which assigns to every possible outcome a value. An agent is then rational if it acts so as to maximize its utility. The associated mathematics of economic and game theory, developed largely since the publication in the 1940s of Von Neumann and Morgenstern's *Games and Economic Behaviour* (Neumann and Morgenstern, 1944) has proved very successful at predicting and understanding a range of economic and social phenomena, and is increasingly finding applications in apparently unconnected disciplines such as evolutionary theory. More recently, the concepts underlying such formalisms have found their way into agent systems (Wellman, 1993).

In contrast, a number of philosophers and social scientists have developed alternative models of rational agency. Philosophers such as Bratman have sought to understand human action through cognitive theories, which attempt to explain how an individual's behaviour emerges through the interaction of mental states such as beliefs, desires, and intentions (Bratman, 1987; Conte and Castelfranchi, 1995). For example, we might say that an individual with an intention to bring about p was being irrational if this individual believed that it was not possible to bring about p (Cohen and Levesque, 1990). Like game and economic theory, such theories have had a significant impact on the

development of the agent field (Rao and Georgeff, 1995). By contrast, social scientists have often tended to see individual rationality as dependent on and subordinate to collective belief systems, and as typically embedded in patterns of normative (and ritual) behaviour (Rappaport, 1984).

The purpose of this panel was to investigate how the concept of rationality is treated from within these different communities, with particular emphasis on the field of agents and multi-agent systems. Each panelist was asked to respond to three questions, dealing with different aspects of rationality and agent systems. The remainder of this article summarizes these questions and the panel responses to them.

2 Bounded rationality

Theories of rational action have, on the whole, failed to deal with the problem of resource bounds*: the fact that any real agent has limited resources (information, computation, memory, time) available in order to make decisions. How do you account for resource boundedness?*

Response by Castelfranchi

Before answering these stimulating questions I absolutely need to introduce a general clarification. Correctly interpreted classical rationality (rational decision theory) should say nothing about goals, motives, or preferences of the agents. It should be just an empty shell, a merely formal or methodological device to decide the best or a satisfying move, given a set of motives/preferences and their importance or order. Thus, being "rational" says nothing about being altruistic or not, being interested in capital (resources, money) or in art or in affects or in approval and reputation! The instrumentalist, merely formal, approach to rationality should not be mixed up with the substantialist view of rationality: instrumentalist rationality ignores the specific motives or preferences of the agents. Thus "utility" should not be conceived as a motive, a goal of the generic agent. Utility is just an abstraction relative to the"mechanism" to choose among the real motives or goals of the agent. Although everybody (especially economists and game theorists) will say that this is obvious and well known, we have to be careful since eventually they are likely to mix up the two things, and, by adopting a rational framework, we will accidentally import a narrow theory of agent's motivation, i.e. the Economic Rationality which is (normative) rationality plus economic motives (profit) and selfishness. Economists and game theorists are the first responsible of such a systematic misunderstanding.

Even adopting a rational decision framework we can postulate in our agents any kind of motive/goal we want or need: benevolence, group concern, altruism, and so on. This does not make them less rational, since rationality is defined *subjectively*. It might make them less efficient, less adaptive, less competitive, less "economically" rational, but not less *subjectively* rational. This distinction—always claimed to be obvious, and yet always ignored—is to me orthogonal to the other distinction between Olympic or perfect or normative rationality, and Simon's limited and bounded rationality: it is not the same distinction.

Even so, "cleaned" decision-theoretic rationality is not necessary (i.e. it is not the only possible device) for rational or adaptive agents (for several reasons, not only because it needs to be bounded).

Turning to the question, my answer is: Simon's theory plus prospect theory, heuristics, dynamic allocation of computational/cognitive resources (by some economics of cognitive resources, or by other techniques), anytime algorithms, and so on. I would also like to add the context dependent activation of goals and of knowledge: the agent should consider/use only the goals, the information, the inferences pertinent to the current context and activated in it. This is not just an unfortunate limitation: it is usually adaptive and efficient. Thanks to the situated activation not all possible profitable investments (activities/goals) are considered, but the choice is only among those agent's goals that are active in that specific situation the agent is involved in. I believe that this situated rationality is quite different from Simon's limited rationality which refers to cognitive limitations and sub-ideal knowledge for rational choice.

Moreover, rationality subordinated to and oriented by the achievement of specific motives (goal-directed or motivated rationality) is not the same, should not make the same prediction and produce the same behaviour than merely formal or instrumentalist rationality, which is oriented by the meta-goal (not a real motive) of maximizing utility. While in the instrumentalist and economic perspective one goal or the other to me is the same, I just chose those goals, I will just allocate my effort and resources in those activities, that promise me the higher profit, in the other perspective, goals are not at all fungible with one another. The agent is interested in specific results (world state), it desires something, it is motivated to achieve a given goal. While in the first perspective the agent will examine all possible goals it knows, and it will follow are source-drive reasoning (''how can I best allocate my resources?''), in the second perspective it starts from goals (motives), it examines only currently active goals (''how can I achieve my goals as much/many as possible?''), and search not for all possible goals but for all possible means and resources for achieve them.

AI should provide models for this difference: it should provide several possible ''rational'' architectures, strategies, and agents, much richer that rational decision theory and *Homo oeconomicus*.

Response by Doran

Having archaeological and anthropological interests, I find it natural to take a long term ''evolutionary'' view of societies. That is, I find it natural and informative to connect particular social phenomena to the long-term survival effectiveness of the society, in its environment, which displays them. Artificial societies are multiple ''pseudo-intelligent'' agent systems within a shared computer-based environment and designed with the intention of studying abstract social processes. They are currently offered as a way to build social theory by means of computer-based experimentation. They have become an important focus of attention, partly just because evolutionary processes may be modelled and understood within them. The artificial society approach to understanding social phenomena contrasts with attempting to build formal logical models, which typically seems obliged to oversimplify in order to achieve even a modest degree of tractability.

Turning to this question specifically, all agents inevitably suffer from bounded rationality. The problems this poses may be summarized as:

- how to capture bounded rationality within a formal logical theory? The hard choice between realism and formal tractability is nowhere more apparent than on this issue.
- how to address the many aspects and implications of bounded rationality in artificial societies—as such this is a (major) part of the general methodological problem of just how to go about studying artificial societies in a systematic and effective way.

Response by Binmore

There is a flourishing literature on what economists call *bounded rationality*. It is expensive to pay attention to things, if only because one could profitably use the computational capacity devoted to this purpose elsewhere. Economists model agents as finite automata and impose costs on the use of more complex automata. A new literature assumes that the automata observed in practice will be determined by an evolutionary process.

3 Reductionism

The sociologist Durkheim suggested the existence of ''social facts''—properties of a social system that could not be explained by examining the individuals within the system, but that could only be viewed as systemic properties. Do you agree? If so, is a theory of individual *rational action going to be sufficient for us to build effective multi-agent systems, or do we need a theory of* social *rationality to deal with such systemic properties? What would such a theory look like? What sorts of predictions would it make?*

Response by Castelfranchi

The problem of the micro-macro link, of reductionism, of the individual-social system relationship, is independent of the rational nature of the agents (see later). This is ''the'' problem of the social sciences according to von Hayek. Only AI—combined with social simulation—can solve this problem: by formally modelling and simulating at the same time the minds of the agents, their behaviours, the emerging systemic phenomena, and their feedback. A theory of emergence is needed (currently there is just a Babel of ill-defined notions), including a theory of ''cognitive emergence'' (something molecules and insects cannot have): the agents becoming partially aware of and modelling the collective effects. A theory of ''immergence'' is also needed: how the emerging phenomenon feeds back into the micro-level, and modifies and shapes the minds and the behavior of the agents, reproducing itself or producing a new emergent phenomenon. There is a co-evolutionary, dialectic relation between micro- and macro- levels that is waiting for some clear theory. For sure cognition is not enough: a lot of social phenomena and cooperation happens unconsciously and unintentionally, (but) very efficiently, also among cognitive intentional agents.

Response by Doran

The artificial societies (models) we create are necessarily conceived in terms of our own conceptual repertoire. It is often convenient to think in terms of the ''level of specification of the society'' and its conceptual repertoire, and then of a further conceptual repertoire needed to express what we observe to be the system behaviour ''emergent'' at higher levels. We, as observers or designers, need to deploy concepts appropriate to a particular level of consideration. But in the system itself (e.g. a computational process on a computer), each successive level (of consideration) may well be fully determined by that below.

Certainly a theory of individual rationality is insufficient. But a theory of social rationality seems to have built into it an elusive and arguably inappropriate ethnocentric concept of rationality. We need a theory of societies that does not emphasize rationality *a priori*. That might look like a set of true statements expressing and predicting systemic properties and heuristically defining and using such concepts as ''agent'', ''role'', ''norm'', ''collective ideology'' and ''emotional energy'' in order to do so.

But the requisite set of concepts should not be prejudged. Building social theory by means of experimentation with artificial societies surely implies not predetermining a particular conceptual repertoire, but discovering it.

Response by Binmore

I believe Durkheim was hopelessly wrong. First, because evolutionary psychology has refuted the *tabula rasa* theory so dear to sociologists. Secondly, because the systemic properties he emphasizes are best modelled as conventions for coordinating on one of the many equilibria of a society's Game of Life. But one cannot even say what an equilibrium *is* without a theory of individual action as a foundation.

4 Society and social reasoning

''There is no such thing as society. There are only individuals.''

Margaret Thatcher.

''We don't live alone. We are members of one body. We are responsible for each other.''

J. B. Priestley (An Inspector Calls)

Margaret Thatcher's famous quote indicates that she believes agents need not reason about society: efficient global behaviour will emerge if agents act in their own self interest, with the barest minimum of external constraint.

J. B. Priestley would not have agreed with Thatcher's views. He argued that we should recognize that we are part of a society, and act in the best interests of this society—individuals should explicitly reason about social structures, and act not only in their own best interests, but in the best interests of the society to which they belong.

Who is right? What do you have to say about each viewpoint? If you agree with Priestley, then how can you reconcile this with individual rational action?

Response by Castelfranchi

Although I sympathize with the socialist claim, as they are formulated they are both wrong, since they both are just prescriptive and ideological. Anyway, how to reconcile individual rationality and group achievements? Given goal-autonomous agents, basically there are two solutions to this problem of making the agent "sensible" to the collective interest:

1. to use external incentives: prizes, punishment, redistribution of incomes, in general rewards, for example money (for example to make industries sensible to the environmental problem you can put taxes on pollution), so that the agent will find convenient—relatively to his/her selfish motives and utility—to do something for the group (to favour the group or to do as requested by the group).
2. to endow the agent with pro-social motives and attitudes (sympathy, group identity, altruism, etc.) either based on social emotions or not, either acquired (learning, socialization) or inborn (by inheritance or design); in this case there is an intrinsic pro-group motivation. The agent is subjectively rational—although not economically rational—but ready to sacrifice.

Human societies use both these approaches; this is not causal. We should experiment advantages and disadvantages of the two, and on which domain and why one is better than the other.

Response by Doran

Trivially agents, including software agents, can and do have and use internal representations of the society of which they are an element, and of its environment. Of course, the representations may be simplistic, partial, and will often be partly or wholly inaccurate.

I argue that the nature of the representations that an agent "should" have is necessarily determined by (a) the other properties of the agents and of their collective "physical" environment and (b) what system properties are to be maximized (e.g. the individual agents well-being in some defined sense over some specified period, or that of the society as a whole). Robots cooperatively collecting pucks are in a very difference situation from humans trying to live "happily"in some defined sense. Viewed as a closed system, the design choice of the agent's internal representations may, in principle, and all else fixed, be used to drive the system into a selected (by the experimenter) state. "Individual rational action" is, or may be, merely one element of the system.

So is it the case that "we should recognize that we are part of a society and act in the best interests of this society" [Priestley]? An answer follows from the preceding paragraph. Assuming an adequate theory of ourselves and our environment (which of course we don't have!) then the answer would straightforwardly depend upon what we are trying to achieve and who exactly "we" are that is trying to achieve it. But, of course, it also follows that we, if inside the system, do not have a free choice but are ourselves determined.

Response by Binmore

Both Thatcher and Priestley are wrong. Priestley is wrong because a workable society cannot rely on people being saints. It has to accept that people will respond to their incentives. Only equilibria in

the Game of Life are therefore viable in the long run as social contracts. Thatcher is wrong because she fails to see that a society *is* defined by the equilibrium its historical experience has taught it to operate.

References

Binmore K, 1992. *Fun and Games: A Text on Game Theory* D. C. Heath and Co.
Bratman ME, 1987. *Intentions, Plans, and Practical Reason* Harvard University Press.
Cohen PR and Levesque HJ, 1990. "Intention is choice with commitment" *Artificial Intelligence* **42** 213–261.
Conte R and Castelfranchi C, 1995. *Cognitive and Social Action* UCL Press.
Neumann JV and Morgenstern O, 1944. *Theory of Games and Economic Behaviour* Princeton University Press.
Rao AS and Georgeff M, 1995. "BDI Agents: from theory to practice" In: *Proc. First International Conference on Multi-Agent Systems (ICMAS-95)* 312–319, San Francisco, CA.
Rappaport RA, 1984. *Pigs for the Ancestors (2nd ed)*. Yale University Press.
Russell S. and Norvig P, 1995. *Artificial Intelligence: A Modern Approach* Prentice-Hall.
Simon HA, 1981. *The Sciences of the Artificial (2nd ed)*. MIT Press.
Wellman MP, 1993. "A market-oriented programming environment and its applications to multicommodity flow problems" *Journal of AI Research* **1** 1–22.