# When are robots intelligent autonomous agents?

Luc Steels

Artificial Intelligence Laboratory
Vrije Universiteit Brussel
Pleinlaan 2, B-1050 Brussels, Belgium
E-mail: steels@arti.vub.ac.be

**Abstract.** The paper explores a biologically inspired definition of intelligent autonomous agents. Intelligence is related to whether behavior of a system contributes to its self-maintenance. Behavior becomes more intelligent (or copes with more ecological pressures) when it is capable to create and use representations. The notion of representation should not be restricted to formal expressions with a truth-theoretic semantics. The dynamics at various levels of intelligent systems plays an essential role in forming representations.

Keywords: intelligence, self-organisation, representation, complex dynamical systems.

## 1 Introduction

What exactly are intelligent autonomous agents? Unless we have some good criteria that are clear targets for the field, it will both be difficult to judge whether we have achieved our aims or to set intermediate milestones to measure whether progress has been made.

The goal of this paper is to present a definition of intelligent autonomous agents. The definition has taken its inspiration from biology (in particular [22], [7]) and is different from traditional definitions currently used in AI, such as the definition based on the Turing test. Our definition is quite tough and no robots can at this point be said to be intelligent or autonomous.

## 2 Agents

Let me start with the notion of an agent. First of all, an agent is a system. This means a set of elements which have a particular relation among themselves and with the environment. Agents need not necessarily be physically instantiated. They could for example take the form of a computer program (a software agent) or a collection of individuals which have common objects (a nation acting as an agent). In this context we are specifically interested in physically embodied agents, as in the case of robotic agents or animals.

Second, an agent performs a particular function for another agent or system. This however makes agents not yet different from other kinds of devices or computer programs. The nature of an agent becomes apparent when we look at the common sense usage of the word. A travel agent for example is not only performing a particular function for us. The travel agent also has a self-interest: It will perform the function if it gets in turn resources to continue its own existence.

So we get a third essential property: an agent is a system that is capable to maintain itself. An agent therefore must worry about two things: (i) performing the function (or set of functions) that determines its role in larger units and thus gives it resources and (ii) maintain its own viability.

This definition of an agent is so far almost completely identical with that of a living system. Living systems are usually defined as systems which actively maintain themselves. They use essentially two mechanisms which suggests that agents need the same:

+ They continuously replace their components and that way secure existence in the face of unreliable or short-lived components. The individual components of the system therefore do not matter, only the roles they play.
+ The system as a whole adapts/evolves to remain viable even if the environment changes, which is bound to happen.

The word agent is currently used in a much more loose way then above. Without a biological perspective it is however difficult to distinguish between agents and other types of machines or software systems. In our view, the term agent is used inappropriately for software agents when they do not have any self-interest whatsoever (as is for example the case in [17]), or when the notion of agent is restricted to a unit that is capable to engage in particular communications [11].

The drive towards self-maintenance is found in biology at many different levels and equivalent levels can be defined for robotic agents:

*The genetic level.* This is the level which maintains the survivability of the species. Mechanisms of copying, mutation, and recombination together with selection pressures operating on the organisms carrying the genes, are the main mechanisms in which a coherent gene pool maintains itself and adapts itself to changing circumstances. For artificial robotic agents, the building plans, the design principles, and the initial structures of one type of agent when it starts its operation correspond to a kind of genetic level. Several researchers have begun to make this level explicit and perform experiments in genetic evolution [9], [**?**].

*The structural level.* This is the level of the components and processes making up the individual agents: cells, cell assemblies, organs, etc. Each of these components has its own defense mechanisms, renewal mechanisms,

and adaptive processes. In the case of the brain, there are neurons, networks of neurons, neural assemblies, regions with particular functions, etc. In artificial systems, they involve internal quantities, electronic and computational processes, behavior systems regulating relations between sensory states and actuator states, etc., but we should probably seem them more as dynamic, evolving entities instead of fixed components [**?**].

*The individual level.* This is the level of the individual agent which has to maintain itself by behaving appropriately in a given environment. In many biological systems (for example bacteria or ant colonies) individuals have no or little self-interest. But it is clear that the individual becomes gradually more important as evolution proceeded its path towards more complexity, and conflicts arise between genetic pressures, group pressures, and the tendency of the individual to maintain itself. Greater individuality seems to be linked tightly with the development of intelligence. In the case of artificial systems, the individual level corresponds to the level of the robotic agent as a whole which has to survive within its ecological niche.

*The group level.* This is the level where groups of individuals together form a coherent whole and maintain themselves as a group. This may include defense mechanisms, social differentiation according to the needs of the group, etc. In the case of artificial systems, the group level becomes relevant when there are groups of robotic agents which have to cooperate in order to survive within a particular ecosystem and accomplish tasks together [18].

For a long time, the natural sciences have made progress by reducing the complexity at one level by looking at the underlying components. Behavior at a particular level is explained by clarifying the behavior of the components at the next level down. For example, properties of chemical reactions are explained (and thus predicted) by the properties of the molecules engaged in the reactions, the properties of the molecules are explained in terms of atoms, the properties of atoms in terms of elementary particles, etc. Also in the case of intelligence, we see that many researchers hope that an understanding of intelligence will come from understanding the behavior of the underlying components. For example, most neurophysiologists believe that a theory of intelligence will result from understanding the behavior of neural networks in the brain. Some physicists go even so far as to claim that only a reduction of the biochemical structures and processes in the brain to the quantum level will provide an explanation of intelligence ([26]).

At the moment there is however a strong opposing tendency in the basic sciences to take a wholistic point of view [6]. This means that it is now understood that there are properties at each level which cannot be reduced to the level below, but follow from the dynamics at that level, and from interactions (resonances) between the dynamics of the different levels ([25]). This suggests that it will not be possible to understand intelligent autonomous agents by

only focusing on the structures and processes causally determining observable behavior. Part of the explanation will have to come from the dynamics in interaction with the structures and processes in the environment, and the coupling between the different levels. A concrete example of this approach is discussed in another paper contained in the Trento Advanced Study Institute proceedings [**?**].

## 3 Autonomy

The term autonomy means essentially 'be relatively independent of'. Thus one can have energy autonomy, in the sense that the robot has on-board batteries so that it is at least for a while independent for its supply of energy. One can also have control autonomy or automaticity. The agent then has a way to sense aspects of the environment and a way to act upon the environment, for example change its own position or manipulate objects. Automaticity is a property that we find in many machines today, for example in systems that control the central heating in a house, or in an airplane that flies in automatic mode.

But usually the concept of autonomy in the context of biology and hence for intelligent autonomous agents when viewed from a biological perspective goes further. Tim Smithers (personal communication, 1992) characterises autonomy as follows:

"The central idea in the concept of autonomy is identified in the etymology of the term: *autos* (self) and *nomos* (rule or law). It was first applied to the Greek city states whose citizens made their own laws, as opposed to living according to those of an external governing power. It is useful to contrast autonomy with the concept of automatic systems. The meaning of automatic comes from the etymology of the term *cybernetic*, which derives from the Greek for *self-steering*. In other words, automatic systems are self-regulating, but they do not make the laws that their regulatory activities seek to satisfy. These are given to them, or built into them. They steer themselves along a given path, correcting and compensating for the effects of external perturbation and disturbances as they go. Autonomous systems, on the other hand, are systems that develop, for themselves, the laws and strategies according to which they regulate their behaviour: they are *self-governing* as well as self-regulating. They determine the paths they follow as well as steer along them."

This description captures the essential point. To be autonomous you must first be automatic. This means that you must be able to operate in an environment, sense this environment and impact it in ways that are beneficial to yourself and to the tasks that are crucial to your further existence. But autonomy goes beyond automaticity, because it also supposes that the basis of self-steering originates (at least partly) in the agent's own capacity to form and adapt its principles of behavior. Moreover the process of building

up or adapting competence is something that takes place, *while the agent is operating in the environment.* It is not the case that the agent has the time to study a large number of examples or to think deeply about how it could cope with unforeseen circumstances. Instead, it must continuously act and respond in order to survive. As Smithers puts it:

"The problem of autonomous systems is to understand how they develop and modify the principles by which they regulate their behaviour while becoming and remaining viable as task achieving systems in complex dynamical environments." (Smithers, personal communication, september 1992).

AI systems built using the classical approach are not autonomous, although they are automatic. Knowledge has been extracted from experts and put into the system explicitly. But the extraction and formalisation has been done by analysts. The original knowledge has been developed by experts. It is not done by the system itself. Current robotic systems are also automatic - but so far not autonomous. For example, algorithms for visual processing have been identified *in advance* by designers and explicitly coded in the computer. Control programs have been invented based on a prior analysis of the possible situations that could be encountered. The resulting systems can solve an infinite set of problems, just like a numerical computer program can solve an infinite number of calculation problems. But these systems can never step outside the boundaries of what was foreseen by the designers because they cannot change their own behavior in a fundamental way.

We see again strong parallels with biology. Biological systems are autonomous. Their structure is not built up by an outside agency, but they develop and maintain their internal structure and functioning through mechanisms like self-organisation, evolution, adaptation, and learning; and they do so while remaining viable in the environments in which they operate.

Another way to characterise autonomy takes the viewpoint of the observer. The ethologist David McFarland, points out that an automatic system is something of which you can fully predict the behavior as soon as you know its internal basis of decision making. An autonomous system on the other hand is a system 'which makes up its own mind'. It is not clear, even not to the original designer, how a system will respond because it has precisely been set up so that responses evolve and change to cope with novel situations. Consequently autonomous systems cannot be controlled the same way that automatic systems can be controlled:

"Autonomous agents are *self controlling* as opposed to being under the control of an outside agent. To be self-controlling, the agent must have relevant self-knowledge and motivation, since they are the prerequisites of a controller. In other words, an autonomous agent must *know* what to do to exercise control, and must *want* to exercise control in one way and not in another way." [22], p.4.

Also in this sense, classical AI systems and current robots are not autonomous because they do not have their own proper objectives and motiva-

tions, only those of their designers.

## 4 Intelligence

AI has wrestled since the beginning with the question of what intelligence is, which explains the controversies around the achievements of AI. Let us first look at some typical definitions and then build further upon the biologically oriented definitions discussed in the previous paragraphs.

The first set of definitions is in terms of comparative performance with respect to human intelligence. The most famous instance of such a definition is the Turing test. Turing imagined interaction with either a human or an intelligent computer program through a terminal. When the program managed to trick the experimenter into believing that it was human often enough, it would qualify as artificial intelligence.

If we consider more restricted versions of the Turing test, for example compare performance of chess programs with human performance, then an honest observer must by now agree that computer programs have reached levels of competence comparable to human intelligence. The problem is that it seems possible (given enough technological effort) to build highly complex programs which are indistinguishable in performance from human intelligence for a specific area, but these programs do not capture the evolution, nor the embedded (contextual) nature of intelligence [4]. As a consequence 'intelligent' programs are often qualified as being no longer intelligent as soon as the person inspecting the program figures out how the problem has been solved. For example, chess programs carry out relatively deep searches in the search space and the impressive performance is therefore no longer thought to be due to intelligence. To find a firmer foundation it seems necessary to look for a definition of intelligence which is not related to subjective judgement.

The second set of definitions is in terms of knowledge and intensionality. For example, Newell has worked out the notion of a knowledge level description [24]. Such a description can be made of a system if its behavior is most coherently described in terms of the possession of knowledge and the application of this knowledge (principle of rationality). A system is defined to be intelligent if a knowledge-level description can be made of it and if it maximally uses the knowledge that it has in a given situation. It follows that artificial intelligence is (almost by definition) concerned with the extraction of knowledge and the formalisation and encoding in computer systems. This approach appears problematic from two points of view. First of all knowledge level descriptions can be made of many objects (such as thermostats) where the label 'intelligence' does not naturally apply. Second, the approach assumes a sharp discontinuum between intelligent and non-intelligent systems and hence does not help to explain how intelligence may have arisen in physical systems nor how knowledge and reasoning relates to neurophysiology.

There are still other definitions, which however are not used within AI

itself. For example, several authors, most notably Roger Penrose, claim that intelligence is intimately tied up with consciousness and self-consciousness [26]. This in turn is defined as the capability to intuit mathematical truths or perform esthetic judgements. The topic of consciousness is so far not at the center of discussion in AI and no claims have ever been made that artificial intelligence systems exhibit consciousness (although see the discussion in [34]). Whether this means, as Penrose suggests, that consciousness falls outside the scope of artificial systems, is another matter. In any case it seems that the coupling of intelligence with consciousness unnecessarily restricts the scope of intelligent systems.

These various definitions are all to some extent controversial. So let me now attempt another approach, building on the definitions so far. This means that intelligence is seen as a property of autonomous agents: systems that have to maintain themselves and build up or adapt their own structure and functioning while remaining viable. But many researchers would argue, rightfully, that intelligence involves more than survivability. The appropriate metabolism, a powerful immune system, etc., are also critical to the survival of organisms (in the case of artificial systems the equivalent is the life time of the batteries, the reliability of microprocessors, the physical robustness of the body). They would also argue that many biological systems (like certain types of fungi) would then be more intelligent than humans because they manage to survive for much longer periods of time. So we need to sharpen the definition of intelligence by considering what kind of functionalities intelligent systems use to achieve viability.

Here we quickly arrive at the notion of representation. The term representation is used in its broadest possible sense here. Representations are physical structures (for example electro-chemical states) which have correlations with aspects of the environment and thus have a predictive power for the system. These correlations are maintained by processes which are themselves quite complex and indirect, for example sensors or actuators which act as transducers of energy of one form into energy of another form. Representations support processes that in turn influence behavior. What makes representations unique is that processes operating over representations can have their own dynamics independently of the dynamics of the world that they represent.

This point can be illustrated with a comparison between two control systems. Both systems have to open a valve when the temperature goes beyond a critical value. One system consists of a metal rod which expands when the temperature goes up and thereby pushes the valve open. When the temperature goes down the metal rod shrinks and the valve closes. There is no representation involved here. The control function is implemented completely in terms of physical processes. The other system consists of a temperature sensor which converts the temperature into a representation of temperature. A control process, for example running on a computer but it could also be an

analogical process, decides when the valve should open and triggers a motor connected to the valve. In this case, there is a clear internal representation and consequently a process operating on the representation which can be flexibly adapted.

From the viewpoint of an external observer there is no difference. Differences only show up when the conditions change. For example, when the weight of the valve increases or when the valve should remain closed under certain conditions which are different from temperature, then a new metal for the rod will have to be chosen or the system will have to be redesigned. When there is a representation, the process operating over the representation will have to change.

Although it seems obvious that the ability to handle representations is the most distinguishing characteristic of intelligent systems, this has lately become a controversial point. Autonomous agents researchers have been arguing 'against representations'. For example, Brooks [3] has claimed that intelligence can be realised without representations. Others have argued that non-representational control systems like the Watt governor are adequate models of cognition [35]. Researchers in situated cognition [4], [27] and in 'constructivist' cognitive science [19] have argued that representations do not play the important role that is traditionally assigned to them. Researchers in neural networks in general reject 'symbolic representations' in favor of sub-symbolic or non-symbolic processing [31]. All this is resulting in a strong debate of representationalists vs. non-representationalists [10]. Let me attempt to clarify the issues.

In classical AI, physical structures acting as representations are usually called symbols and the processes operating over them are called symbol processing operations. In addition the symbol processing is subjected to strong constraints: Symbols need to be defined using a formal system and symbolic expressions need to have a strict correspondence to the objects they represent in the sense of Tarskian truth-theoretic semantics. The operations that can be performed to obtain predictive power must be truth-preserving.

These restrictions on representations are obviously too narrow. States in dynamical systems [14] may also behave as representations. Representations should not be restricted to those amenable to formal semantics nor should processing be restricted to logically justified inferences. The relation between representations and reality can and usually is very undisciplined, partly due to the problem of maintaining a strict correspondence between the environment and the representation. For example, it is known that the signals received by sonar sensors are only for 20 percent effectively due to reflection from objects. Sonar sensors therefore do not function directly as object detectors and they do not produce a 'clean representation' of whether there is an object or not in the environment. Rather they establish a (weak) correlation between external states (the presence of obstacles in the environment) and internal states (hypothesised positions of obstacles in an analogical map) which may

be usefully exploited by the behavioral models.

Second, classical AI restricts itself mostly to *explicit representations*. A representation in general is a structure which has an influence on behavior. Explicit representations enact this influence by categorising concepts of the reality concerned and by deriving descriptions of future states of reality. An implicit (or emergent) representation occurs when an agent has a particular behavior which is appropriate with respect to the motivations and action patterns of other agents and the environment but there is no model. The appropriate behavior is for example due to an historical evolution which has selected for the behavior. The implicit representations are still grounded in explicit representations but these are at a different level. Implicit representations are much more common than is thought, and this, it seems to me, is the real lesson of "situated cognition". So far most successes of classical AI are based on explicit representations which have been put in by designers (and are therefore not autonomously derived).

## 5 Conclusions

The paper discussed a characterisation of intelligent autonomous agents which finds its inspiration in biological theory. It starts from the idea that agents are self-sustaining systems which perform a function for others and thus get the resources to maintain themselves. But because they have to worry about their own survival they need to be autonomous, both in the sense of self-governing and of having their own motivations. Because environments and users of systems continuously change, agents have to be adaptive. Intelligence helps because it gives systems the capacity to adapt more rapidly to environmental changes or to handle much more complex functions by bringing in representations. Intelligence is seen at many different levels and is partly due to the coupling between the different levels. Representations are not necessarily explicit but may be implicit.

Although much progress has been made on many aspects, it is at the same time clear that truly intelligent autonomous agents do not exist today and it will be quite a while before such systems come into existence. Impressive results have been obtained in classical AI using complex representations but these representations have been supplied by designers and are not grounded in reality. Impressive results have also been obtained with classical control systems but such systems hardly use complex representations. Moreover there is no tradition for viewing robots or software systems as living entities that are themselves responsible for their survival.

## 6 Acknowledgement

The viewpoints discussed in this paper have been shaped and greatly enhanced by discussions with many people, including discussions at the Trento

NATO ASI. Thanks are due in particular to Bill Clancey, Thomas Christaller, David McFarland, Rolf Pfeifer, Tim Smithers, and Walter Van de Velde. This research was partially sponsored by the Esprit basic research project SUBSYM and the DPWB concerted action (IUAP) CONSTRUCT of the Belgian government.

## References

1. Arkin, R. (1989) Motor Schema based mobile robot navigation. Int. Journal of Robotics Research. Vol 8, 4 p. 92-112.
2. Brooks, R. 1991. Intelligence without reason, IJCAI-91, Sydney, Australia, pp 569–595.
3. Brooks, R. 1991. Intelligence without representation, AI Journal.
4. Clancey, W.J. (1993) Situated Action: A Neuropsychological Interpretation. Cognitive Science 17(1), 87-116.
5. Cliff, D., I. Harvey, and P. Husbands. (1993) Explorations in evolutionary robotics. Adaptive Behavior 2(1), 71-104.
6. Cohen, J. and I. Stewart (1994) The Collapse of Chaos.
7. Dawkins, R. (1976) The Selfish Gene. Oxford University Press. Oxford.
8. Engels, C. and G. Schoener (1994) Dynamic fields endow behavior-based robots with representations. Robotics and Autonomous Systems. 1994.
9. Floreano, D. and F. Mondada (1994) Automatic Creation of an Autonomous Agent: Genetic Evolution of a Neural-Network Driven Robot. In: Cliff, D. et.al. (1994) From animals to animats 3. Proceedings of the 3d International Conference on Simulation of Adaptive Behavior. MIT Press. Cambridge Ma. p. 421-430.
10. Hayes, P. K. Ford, and N. Agnew (1994) On Babies and Bathwater. A Cautionary Tale. AI Magazine. 15 (4), 15-26.
11. Genesereth, M. and S. Ketchpel. (1994) Software Agents. Comm. of the ACM. 37(7). p. 48-53.
12. Genesereth, M. and N. Nilsson (1987) Logical Foundations of Artificial Intelligence. Morgan Kaufmann Pub. Los Altos.
13. Haken, H. (1983). *Advanced synergetics: instability hierarchies of self-organisating systems and devices*, Springer, Berlin.
14. Jaeger, H. (1994) *Dynamic Symbol Systems* Ph.D. thesis. Faculty of Technology. Bielefeld.
15. Kaneko, K. (1994) Relevance of dynamic clustering to biological networks. Physica D 75, 55-73.
16. Kiss, G. (1993) Autonomous Agents, AI and Chaos Theory. In: Meyer, J.A., et.al. (eds.) *From Animals to Animats 2* Proceedings of the Second Int. Conference on Simulation of Adaptive Behavior. MIT Press, Cambridge. pp. 518-524.
17. Maes, P. (1994) Agents that Reduce Work and Information Overload. Comm. of the ACM 37(7). p. 30-40.
18. Mataric, M. (1994) Learning to Behave Socially. In: Cliff, D. et.al. (1994) From animals to animats 3. Proceedings of the 3d International Conference on Simulation of Adaptive Behavior. MIT Press. Cambridge Ma. p. 453-462.

19. Maturana, H.R. and F.J. Varela (1987) The Tree of Knowledge: The Biological roots of Human Understanding. Shamhala Press, Boston.
20. McFarland, D. (1990) Animal behaviour. Oxford University Press, Oxford.
21. McFarland, D. and T. Boesser (1994) Intelligent Behavior in Animals and Robots. MIT Press/Bradford Books, Cambridge Ma.
22. McFarland, D. (1994) Towards Robot Cooperation. Proceedings of the Simulation of Adaptive Behavior Conference. Brighton. MIT Press.
23. McFarland, D., E. Spier, and P. Stuer (1994)
24. Newell,A. (1981). The knowledge level, *Journal of Artificial Intelligence*, vol 18, no 1, pp 87–127.
25. Nicolis, G. and I. Prigogine (1985) Exploring Complexity. Piper, Munchen.
26. Penrose, R. (1990) The Emperor's New Mind. Oxford University Press. Oxford.
27. Pfeifer, R. and P. Verschure (1992) Distributed Adaptive Control: A Paradigm for Designing Autonomous Agents. In: Varela, F.J. and P. Bourgine (eds.) (1992) Toward a Practice of Autonomous Systems. Proceedings of the First European Conference on Artificial Life. MIT Press/Bradford Books, Cambridge Ma. p. 21-30.
28. Schöner, G. and M. Dose (1993) A dynamical systems approach to task-level system integration used to plan and control autonomous vehicle motion, *Journal of Robotics and Autonomous Systems*, vol 10, pp 253–267.
29. Simon, H. (1969) The Sciences of the Artificial. MIT Press, Cambridge Ma.
30. Smithers, T. (1994) Are autonomous agents information processing systems? In: Steels, L. and R. Brooks, (Eds.), *The 'artificial life' route to 'artifical intelligence': building situated embodied agents*, Lawrence Erlbaum Associates, New Haven.
31. Smolensky, P. (1986) Information Processing in Dynamical Systems. Foundations of Harmony Theory. In: Rumelhart, D.E., J.L. McClelland, (eds.) *Parallel Distributed Processing. Explorations in the Microstructure of Cognition.* Vol 1. MIT Press, Cambridge Ma. pp. 194-281.
32. Steels, L. (1994a) The artificial life roots of artificial intelligence. Artificial Life Journal, Vol 1,1. MIT Press, Cambridge.
33. Steels, L. (1994b) A case study in the behavior-oriented design of autonomous agents. Proceedings of the Simulation of Adaptive Behavior Conference. Brighton. MIT Press. Cambridge.
34. Trautteur, G. (ed.) (1994) Approaches to consciousness. Kluwer Academic Publishing. Amsterdam.
35. Van Gelder, T. (1992) What might cognition be if not computation. Dept of Cognitive Science. Indiana University, Bloomington. Technical report.
36. Winston, P. (1992) Artificial Intelligence. Addison-Wesley Pub. Cy. Reading Ma.

This article was processed using the LaTeX macro package with LLNCS style