Cogn Comput
DOI 10.1007/s12559-012-9129-4

Autonomous Cognitive Systems in Real-World Environments: Less Control, More Flexibility and Better Interaction

Vincent C. Müller

Received: 15 November 2011 / Accepted: 16 February 2012

Abstract In October 2011, the "2nd European Network for Cognitive Systems, Robotics and Interaction", EU-CogII, held its meeting in Groningen on "Autonomous activity in real-world environments", organized by Tjeerd Andringa and myself. This is a brief personal report on why we thought autonomy in real-world environments is central for cognitive systems research and what I think I learned about it. The theses that crystallized are that (a) autonomy is a relative property and a matter of degree, (b) increasing autonomy of an artificial system from its makers and users is a necessary feature of increasingly intelligent systems that can deal with the real world and (c) more such autonomy means less control but at the same time improved interaction with the system.

Keywords Cognitive system · Autonomy · Flexibility · Human–computer interaction · HCI

Cognitive Systems: Flexibility for Intelligence

As the conference organizers, we formulated the theme of the event as follows: "We aim to build autonomous artificial cognitive systems that are to pursue their goals successfully in real-world environments that cannot be fully anticipated, that are not fully known and that change continuously, including other agents …" [1]. This formulation is based on the view that the distinguishing feature of 'cognitive' artificial systems is not so much the use of higher level, traditionally called 'cognitive' features, but rather flexibility in pursuing the goals of the system—and one way to achieve this flexibility is through higher level features, such as planning. (the formulation is inspired by the mission statement of the EU Unit "Cognitive Systems and Robotics" on http://cordis.europa.eu/fp7/ict/cognition).

Traditional programming, including traditional AI programming, tries to anticipate the perceptual input the system might face from environments and to provide a response to this input. Since sensors are not perfect, actual environments cannot be fully sensed and not be fully described, they are often approximated by probabilistic systems—but the basic setup is the same: we try to provide an algorithmic connection between input and output. "And what else could we possibly do?" you might ask. Well, we could provide the whole system with features that will make it achieve its goals, which will not only include bodily features, but also algorithms that make success more probable, even in environments that we have not anticipated.

The extreme case on the one end of the spectrum would be the lookup-table kind of intelligence, which has a finite list of possible states of the environment in the one-half of the table and a specific action (or set of actions) in the other half. This kind of simple reflex system is "pre-programmed" or "scripted". Note, however, that complex combinations of factors in such a system can produce fairly complex behavior that may be 'surprising' to the makers.

One extreme case on a different end of the spectrum is a system that has no information-processing capability, but only responds to its environment due to its other features, e.g., its morphology. In this case, no prior "scripting" takes

V. C. Müller (✉)
Anatolia College/ACT, 55510 Pylaia, Greece
e-mail: vmueller@act.edu
URL: http://www.typos.de

V. C. Müller
Department of Philosophy, FHI, Programme on the Impacts of Future Technology, University of Oxford, Oxford OX1 1PT, UK

Published online: 01 March 2012

place, but the system is equally inflexible. (I assume that information-processing, digital or analog, is non-trivially distinguishable from other causal processes by the presence of components that play the functional role of representation—a direction that is explored, for example, in [2] or [3, ch. 2].)

I think it is important *not* to characterize the feature of flexibility in terms of 'action selection', since this is often a misleading term for the behavior of intelligent systems: The problem "Which action to select next?" stems from the "Model-Plan-Act" view of action (and the "Intention-Belief-Desire" psychology)—the standard in classical representational AI. This view focuses on 'action' that is under voluntary control, which leaves out a large part of intelligent behavior. Many intelligent agents do not seem to 'select actions', be they natural (a slug or even a cockroach) or artificial (non-classical designs, coupled embodied systems; e.g., a passive dynamic walker does not select the action "make the next step").

Autonomy

The demand for flexibility thus results in a demand for autonomy. In a first approximation, autonomy is the ability of the system to respond to the environment "by itself", without a prior "script". The Greek word "autonomy" means the ability to give oneself *(e)avto* ones own law *nomos,* thus becoming *avtonomos*.

However, notions like generating behavior "by itself" or giving laws "to oneself" are of limited use since the system is never alone, but is always subject to various causal influences from the "outside" (if boundaries can even be specified). In control engineering, autonomy is defined in terms of interference by the *user* of the system, where degrees go from a fully user controlled system to a system not controlled at all. (Some systems are adaptive in this respect, i.e., change their level of autonomy depending on needs [cf. 4].) The notion we appear to need is a relational one:

> X is autonomous from Y if and only if X pursues its goals without input from Y

This initial definition can be clarified in at least two ways: The input from Y can be a matter of degree, and thus X can be more or less autonomous from Y. Also, in many cases it will be useful to differentiate that X is autonomous from Y with respect to some type of input or some type of ability (e.g., spatial orientation).

Note that the Y can be a human being, like the maker or user of an artificial system, but we can naturally use the same notion for other agents also: One can say that an animal is more or less autonomous (e.g., the very young are often less autonomous) and that autonomy is reduced with respect to other members of the species, or of other species—e.g., in a plant that lives in symbiosis with another, or in a microorganism that lives inside a larger organism. On the other hand, if a living being is dependent on a specific environment, like a particular cave, for example, I would not call that a reduction of its autonomy. It follows that X can be autonomous from Y properly only if both X and Y are agents.

One view that is often heard is that a system is autonomous to the degree that it can set its own goals. However, a system that "sets its own goals" can do so only within limits provided by the unalterable features of its design, or, to put it differently: there must be causes for any given setting of goals. We say that humans have free will, even though we assume that there are causes for our setting of goals. If humans consider their goals, decide that they want to have different goals and adopt them, then this re-setting is dependent on evaluating the present goals on the basis of higher-order goals. Indeed, this ability to rationally set goals and follow them has been considered the core of one common explanation for what is meant by the idea of human "free will". This approach typically allows for a notion of responsibility in a world where all events are caused [5]. It seems, therefore, that the ability to set goals is just a further degree of autonomy with respect to another system; it is just a form of pursuing higher order goals. Of course, there is the Maturana-Varela tradition, according to which only living autopoetic systems can truly have goals, but is this relevant for our purposes, the understanding of autonomy? Perhaps, it marks an outer bound that certain degrees of autonomy can only be reached by living systems—but I would advise against including this ideology in an initial understanding of the notion (keeping in mind that any system will have several goals, often on several levels of description or granularity).

Given this discussion, we might accept a formulation like this:

> Agent X is autonomous from agent Y to the degree that X pursues its goals without input from Y

Flexibility Through Autonomy in an Environment

If we assume that intelligence is the ability to pursue one's goals successfully, then it becomes clear that more autonomy does not necessarily imply more intelligence. In particular, autonomy will only be a useful feature of intelligence in environments that are not fully anticipated. It is for this reason that in real-world environments, which can only be anticipated to a very small degree, autonomy is a crucial feature of intelligence—in fact *the* crucial feature,

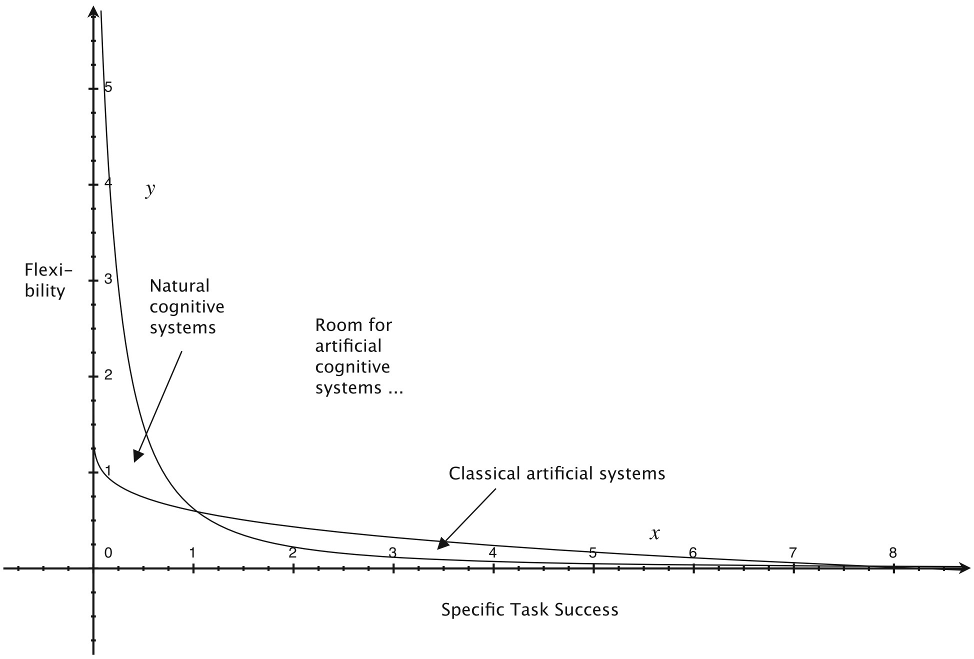

Fig. 1 The complexity space for intelligent agents

together with successful use of the environment, including other agents (and even these features of the environment are not specifiable ahead of time, so they require autonomy, too). In other words, in the real world, an artificial system must be strongly autonomous of its maker and user to be successful or intelligent. This stands in stark contrast to the totally controlled environment of simulation and the highly controlled environments of industrial robotics today. (One major reason why the Fukushima disaster happened and was then dealt with so badly is the absence of autonomous and thus flexible systems in both phases—no current robot can deal with the inside of the reactor if we cannot specify in advance what the inside is like [6].)

One way to explain this point is to look at the agent as carrying out a task of some complexity in an environment of some complexity (Fig. 1).

In this space of possibilities, classical robotics provides solutions in the area that is low in y values, but increasingly more successful in x values—it achieves complex tasks in non-complex environments. Natural systems, on the other hand, are able to deal with high y values, but achieve relatively low scores in the x range—they achieve less complex tasks but in complex environments. Humans are probably an extreme example in terms of their adaptability to different environments while also achieving fairly high scores in certain x sectors—but humans remain the "faulty beings", as Hans Gehlen called them, that are surpassed in nearly any ability by some animal (except higher cognition). Artificial cognitive systems will enable artificial systems to 'detach' themselves from the y baseline and move into the higher spaces opened up by y. This move into more complex unspecified environments requires more autonomy on the part of the systems. (This notion owes much to the discussions in the 2011 EUCogII workshop "Challenges for Cognitive Systems" at Rapperswil.)

Autonomy as Loss of Control Versus Autonomy as Gain of Interaction

Given that we have defined autonomy in terms of independence from other agents, in particular the maker and user of a system, it follows that more autonomy implies less control. This loss of control would seem undesirable both in terms of performance of the system and in terms of avoiding undesired behavior of the system, which would also be less predictable.

I would suggest, however, that these negative effects are outweighed by positive ones, namely the improved ability of humans (and other agents) to interact with such systems. In a different context, I have recently made the obvious point that human interaction with other agents relies crucially on the attribution of intentions or goals to these agents [7]. We humans can interact successfully with other agents just when we can interpret their behavior as driven by their goals, shown by behavior that provides a resistance to our own. It is impossible to cooperate with another agent that has no goals, thus offering no such resistance. Such interaction will also be a lot easier than technical control—since we humans are already experts at it.

I conclude that new human–computer interaction (New HCI) will be properly a form or interaction, less of control. We must overcome the wish to *control* the autonomous cognitive system and begin to *interact* with it.

References

1. Andringa T, Müller VC. 10–11 October 2011—Fifth EUCogII Members Conference, Groningen, “Autonomous activity in real-world environments”. Available from: http://www.eucognition.org/index.php?page=fifth-conference-general-info.
2. Müller VC. Is there a future for AI without representation? Mind Mach. 2007;17(1):101–15.
3. Miłkowski M. Explaining the computational mind. Cambridge: MIT Press (forthcoming).
4. Sheridan TB. Humans and automation: system design and research issues. New York: Wiley; 2002.
5. Frankfurt H. Freedom of the will and the concept of a person. J Philos 1971;LXVIII(1):5–20.
6. Cockburn-Price S, Müller VC. Robots seek role in reactors: artificial cognitive systems will soon be able to play a greater role in dealing with problems such as Fukushima. Professional Engineering 24/5, 04.05.2011. Available from: http://profeng.com/news/robots-seek-role-in-reactors_1083.
7. Müller VC. Interaction and resistance: the recognition of intentions in new human–computer interaction. In: Esposito A, Esposito AM, Martone R, Müller VC, Scarpetta G, editors. Towards autonomous, adaptive, and context-aware multimodal interfaces: theoretical and practical issues. Berlin: Springer; 2011. p. 1–7.