The paradox of autonomy: The interaction between humans and autonomous cognitive artifacts

Alexander Riegler
Center Leo Apostel for Interdisciplinary Research
Vrije Universiteit Brussel
Krijgskundestraat 33, B-1160 Brussels
ariegler@vub.ac.be

Abstract. According to Thrun and others, personal service robots need increasingly more autonomy in order to function in the highly unpredictable company of humans. At the same time, the cognitive processes in artifacts will become increasingly alien to us. This has several reasons: 1. Maturana's concept of structural determinism questions conventional forms of interaction. 2. Considerably different ways of embodiment result in incompatible referential frameworks (worldviews). 3. Engineers focus on the output of artifacts, whereas autonomous cognitive systems seek to control their input state. As a result, instructional interaction – the basic ingredient of conventional man–machine relationships – with genuine autonomous systems will become impossible. Therefore the increase of autonomy will eventually lead to a paradox. Today we are still in a position to anthropomorphically trivialize the behavioral pattern of current robots (von Foerster). Eventually, however, when self-organizing systems will have reached the high levels of autonomy we wished for interacting with them may become impossible since their goals will be completely independent of ours.

Introduction

The success of virtual life forms since the late 1990s shows that humans can develop feelings for autonomous artifacts. Examples of entertainment agents are abundant. On the software level we find Tamagotchi (a tiny cyberspace pet that needs constant care in order to grow bigger and more beautiful), the Sims (a group of virtual people who have a set of 'motives' that need to be satisfied), and Kyoko (the Japanese virtual teen-age idol singer). Implemented in hardware there are 'creatures' such as Sony's AIBO robodog (which understands dozens of voice commands but does not necessarily obey) and Kismet (a robot with humanlike expressiveness and the ability to interact with people like a growing child).

In these examples the interaction between the human audience and the artifact is dominated by feelings of sympathy and acceptance. This observation can be accounted for by the fact that entertainment robots are not supposed to carry out a particular task; they merely act as a pet. In contrast to such toys the interaction with service robots follows a different pattern. They are made for specific ends and are therefore under the control of humans. It is a relationship determined by power which applies to robots and computers alike: "Users are empowered by having a clear predictive model of system performance and a sense of mastery, control, and accomplishment." (Shneiderman in Don et al. 1992). Thrun (2004) distinguishes three types of purposeful robots: 1. industrial robots, 2. professional service robots, and 3. personal service robots. Although these categories mainly refer to their domain of action they also reflect the degree of autonomy the robots are equipped with. While robots in industrial settings are supposed to mechanically repeat a closely defined range of working steps, need

personal robots (such as robotic vacuum cleaners and robodogs) a high degree of autonomy. As Thrun put it, "robots which operate in close proximity to people require a high degree of autonomy, partially because of safety concerns and partially because people are less predictable than most objects." Due to the rising public interest in robots of category 3 engineers are requested to design cognitive artifacts of increasingly more extended autonomy. I claim that this development will lead to a paradox where the interests of artifact designers collide with intrinsic properties of cognitive autonomy.

The remainder of the paper is concerned with first identifying the fundamental differences between humans and cognitive machines. The idea is to make a sharp distinction between behavior-based modeling and the input-centered perspective of the cognitive entity. Due to this fundamental gap, as observer-designers we are driven into trivializing cognitive systems. Consequently, our bias to model even most complex phenomena in terms of input–output relations thwarts our attempts to build genuinely autonomous systems. Finally, I discuss the implications of the autonomy paradox and how it will change our interaction with future generations of cognitive artifacts.

Alien cognition

In his 2002 paper, Alois Knoll succinctly pointed out that "humanoids will never be an exact replica of human intelligence." It can be argued that cognitive artifacts will always be differently embodied than humans. They are composed of inorganic materials and they lack phylogenetic ancestry, whereas "we human beings are the arising present of an evolutionary history in which our ancestors and the medium in which they lived have changed together congruently around the conservation of a manner of living in language, self-consciousness and a family life", as Maturana (2005) put it. Furthermore, artifacts have only a rudimentary ontogeny in the sense that their learning capacity either does not exist or is based on arbitrarily defined or incomprehensible algorithms. Here are four examples.

In what I called *PacMan* systems (which includes artificial life simulations as well as toys like Tamagotchis) anthropomorphically defined entities try to optimize certain parameters which are interpreted as their behavior. In such models it seems arbitrary why a certain pixel on the screen represents 'food' for the creature. Rather, it is the programmer who imposes meaning onto it. In *situational systems* robots merely *cope* with the present context and react to instantaneous information rather than construct sophisticated mental models of their environment. In *symbolic rule-based* systems the artifact has to base its cognition on readily prepared propositions that cut the world into crisp anthropomorphic categories. And while *massively distributed* systems using unsupervised learning algorithms seem to display 'natural' behavior they have a rather poor explanatory performance, which obfuscates the cognitive details. (For a more detailed discussion, cf. Riegler in press).

As a consequence, cognition in artifacts is and will remain alien to us. Chess computers play their game entirely different than humans. The former exhaustively explore all possible consequences of a certain move, the latter rely on intuitive heuristics. Poems prove to be an obstinate problem for translation programs because they are implemented in a way that focuses on formal syntactical patterns. Even the introduction of semantic ontologies does not greatly improve the problem as it merely shifts the problem of arbitrary reference between cognition and token. Consequently, such highly rational and detached systems are unable to capture a human being's referential system indispensable for text interpretation.

It can be argued that the alienation merely depends on choosing the 'proper' algorithm. However, there are reasons to assume that it is based on a fundamental epistemological aspect, which I detail in the following section.

The fundamental gap

I shall argue that the alienation of artificial cognitive systems arises from neglecting the difference between two perspectives, P1 and P2.

P1. Human designers concentrate on the *output* of an artifact: a robotic lawn mower should cut the grass efficiently, a navigational system should find the shortest route to save time and energy, etc. These technological artifacts do have a clearly pre-defined goal. Their autonomy is confined to finding the optimal chain of sub-goals in order to meet their purpose. Maes (1995) described artifacts equipped with this sort of autonomy as "computational systems that inhabit some complex dynamic environment, sense and act autonomously in this environment, and by doing so realize a set of goals or tasks for which they are designed." These are "automatic systems […which] steer themselves along a given path, correcting and compensating for the effects of external perturbation and disturbances as they go" (Smithers, quoted in Steels 1995). Ziemke (1998) referred to their mode of working as "operational autonomy", i.e., "the capacity to operate without human intervention". For purposes of simplicity, let us call this sort of self-steering *subgoal-autonomy*, or *s-autonomy* for short.

From an engineering point of view, operational subgoal autonomy may prove inadequate in complex environments where the original human-specified goal becomes ill-defined, outdated, or inferior to other solutions. Based on Smithers demand that systems should "develop, for themselves, the laws and strategies according to which they regulate their behavior," Ziemke introduces the notion of "behavioral autonomy". However, as argued below, for natural systems and future artifacts this notion may be misleading and should be called *emerging autonomy*, or *e-autonomy*, instead.

P2. In contrast to robot builders, natural animals control their *input* (perception) rather than their output (behavior), or as Powers (1973) put it, "behavior is the process by which organisms control their input sensory data" (cf. also Porr & Wörgötter 2005 who discuss input control in more technical terms). This perspective is based on the concept of homeostasis (Cannon 1932), which says that a living organism has to keep its intrinsic variables within certain limits in order to survive. Such "essential variables" (Ashby 1952) include body temperature, levels of water, minerals and glucose, and similar physiological parameters, as well as other proprioceptively or consciously accessible aspects in higher animals and human beings. Consequently, natural systems execute certain actions in order to control and change their input state, such as avoiding the perception of an obstacle or drinking to quench their thirst. The state of homeostasis is reached by the process of negative feedback. Like in a thermostat, a given parameter is kept under control by appropriate counteractions. As well known, the thermostat adjust the temperature by turning off the heating as soon as a certain temperature is reached and turning it on as soon as the temperature drops below a certain reference value. While such a simple feedback loop may suffice for primitive intrinsic variables, higher order goals are accomplished in a hierarchical assemble of feedback loops in which each level provides the reference value for the next lower level: "The entire hierarchy is organized around a single concept: control by means of adjusting reference-signals for lower-order systems" (Powers 1973). So at higher levels the system controls the output of lower levels, at the bottom, however, its perceptual input.

The anthropomorphizations of behavior

Making the P1–P2 distinction implies that for observers and designers of artifacts the situation is diametrically different from the situation of the modeled artifacts. The former lack the possibility to sense the system's input state (or, philosophically speaking, "first-person experience"), so they focus on the system's output. If they want to understand a living organism and try to model it they will face with the problem of reverse engineering, i.e., the

reconstruction of a functional model (either in the abstract-mathematical or material domain) that copies the original model's *behavior*. This amounts to the fact that the modeler defines systems over the range of their behaviors and builds them accordingly. The result are *anthropomorphic* rather than autonomous artifacts. Consequently, reverse engineering in the cognitive domain means postulating a link between the rule-like behavior of observed of natural cognitive systems and general laws of cognition and knowledge acquisition. However, observers – whether ethologists or engineers – are not embodied in the world of the observed animal (Nagel 1974) nor designed artifact (Riegler 2002). They interpret its behavior within their own referential system of understanding, usually called *worldview*. According to Aerts et al. (1994), a worldview can be described as a system of co-ordinates or a frame of reference in which everything presented to us by our diverse experiences can be placed. Such a representational system allows us to integrate everything we know about the world and ourselves into a global picture, one that illuminates reality as it is presented to us within a certain context. Within this framework, moving and living entities are closely linked with personal states of mind and emotional dispositions, as entertainment agents demonstrate. This bias to interpret the movements of artifacts in a particular way was characterized by Kiesler & Hinds (2004) as follows, "People's mental models of autonomous robots are often more anthropomorphic than are their models of other systems… The tendency for people to anthropomorphize may be fed, in part, by science fiction and, in part, by the powerful impact of autonomous movement on perception."

What follows from the P1–P2 distinction is that there are two domains of description, one which describes the (natural or artificial) system in question in relation with its environment and one that describes the actual working of the system in terms of its components. To paraphrase Maturana (1974), what occurs within a composed entity (e.g., living systems) is different from what happens to this entity. This is particularly obvious in the case of living systems, which exist in the domain of physiology and in the domain of behavior. These two phenomenal domains do not intersect since the description of a composite unity takes place in a meta-domain with respect to the description of the components that constitute that unity. An observer may simultaneously look at both. They might state that changes in physiology cause changes in the behavior of an observed creature, but that is not a logical necessity. Or as Glasersfeld (1979) put it, "First, there is the tempting but logically erroneous idea that what we rightly call 'environment' relative to an organism when both the organism *and* its environment are being observed by us, must also be *our* environment and can, therefore, be held causally responsible for what we ourselves experience. Second, there is the mistaken belief that the 'environment' which is part of our experiential field has to be identical with the experiential field of the observed organism." But "… in order to explain a given behavior of a living system, the observer must explain the generation and establishment of the particular structures of the organism and of the environment that make such behavior possible at the moment it occurs." (Maturana 1974).

In the context of artificial intelligence, it becomes clear why Turing (1950) so vehemently rejected the "Can machines think?" question. His imitation game is nothing more than admitting that there is no other way to assess the cognitive abilities of an artifact than approximating them in an output-oriented test.

Behavioral trivialization

What is the consequence for artificial autonomous systems? As pointed out before, cognitive engineers would need the intellectual capacity of making inferences from behavior to inner working. Such reverse engineering, however, requires checking a large number of possible mappings from observational data onto the model. As known from Duhem underdeterminism theorem and extended by McAllister (2003) there is an arbitrarily large number of ways to

explain data points: “Any given data set can be interpreted as the sum of any conceivable pattern and a certain noise level” (McAllister 1999). This becomes immediately obvious in the case of black boxes whose inner mechanisms are unknown to the outside observer. It is extremely difficult to make inferences. Consider a box with four input, four internal, and four output states, all of which can be wired in any way. The total number of possible configurations is $4^{44} = 2^{32} \approx 4 \times 10^9$. In other words starting from a behavioral protocol one needs to test 4 billion different models to find the one that produced the recorded behavior.

Facing this combinatorial obstacle, all we can do is trivialize systems of this complexity (Foerster 1970) in order to deal with them as cognitive scientists and engineers. This means to reduce the degrees of freedom of a given complex entity in order to behave like a ‘trivial machine’, i.e., an automaton without internal states, which responses to an input with always the same output. Trivialization can be accomplished by anthropomorphically attributing behavior, i.e., by “projecting the image of ourselves into things or functions of things in the outside world” (Foerster 1970). Modeling is therefore the procedure that trivializes complex living systems. Maturana (1974) notes that it possible to “treat an autopoietic [i.e., living] system as if it were an allopoietic [i.e., man-made] one by considering the perturbing agent as input and the changes that the organism undergoes while maintaining its autopoiesis as output. This treatment, however, disregards the organization that defines the organism as a unity by putting it in a context in which a part of it can be defined as an allopoietic subsystem by specifying in it input and output relations.”

So there are two non-overlapping domains, the actual working of the robots as *e*-autonomous system and the conceptual domain of the designer-user who would like the artifact to perform certain tasks in an *s*-autonomous way. This is the autonomy paradox.

The paradox makes it appear unlikely that our interaction with *e*-autonomous artifacts will be one of control and mastery. Such instructive interactions are impossible because the instructions *e*-autonomous systems undergo “will only trigger changes in them; they will not specify what happens to them” (Maturana 1987).

Domestication instead of interaction?

Although the autonomy paradox has implications on how we will interact with future artifacts it is not the first time in history that humans would learn to cope with autonomous systems. Thousands of years ago people started to turn wild beasts into domesticated pets. The domestication of highly autonomous and partly unpredictable wild animals must be considered a case of trivialization. No longer does the horse freely roam the steppe. It is now used a transport vehicle with the clearly defined goal of moving from A to B. In other words the domestication of natural autonomous systems is an indication that despite the autonomy paradox direct and instructive interaction with sophisticated autonomous artifacts may be possible.

However, there is a caveat. As pointed out by Diamond (2002), only 14 out of 148 mammals could be successfully domesticated. In nine out of ten cases domestication failed as not all of the following criteria could be met. These criteria can be divided into two physiological and four cognitive conditions. (P1) There must be an easy and inexpensive way to supply the animals with food. This means that carnivores and other inflexible eaters such as lions are not eligible. (P2) Animals must have a reasonably fast growth rate and short birth intervals. This is the reason why elephants got tamed but never domesticated. (C1) Taming refers to the fact that the individual animal is bred in the wild and captured afterwards. Species that are generally reluctant to breed in captivity are therefore poor candidates for domestications. The lama-like vicuñas in South America display this shortcoming, which makes it impossible to domesticate these animals whose fur is highly appreciated and priced. (C2) Creatures must have a pleasant disposition unlike grizzly bears, which have delicious

meat, or zebras, which could have been used as an alternative to horses. (C3) Animals must no show the tendency to panic easily. Deer and gazelles are therefore excluded from domestication. (C4) Animals must live in modifiable social hierarchies which are dominated by a leader and which can be penetrated by a human. Antelopes do not have such a social structure, and neither have cats. This is the reason why there are no domesticated cats; they do not obey commands.

The list of criteria shows that it might turn out rather difficult to use highly autonomous systems for anthropomorphic goals. In the case of animals, the required six conditions cover both physiological and cognitive areas. Future research will have to pint out how the six conditions map onto artifacts. Among the cognitive conditions, C2 could be interpreted as the three robots laws as formulated in Isaac Asimov's novels. It clearly assigns priority to the protection of human beings. C3 may refer to a solid action-selection algorithm taking care that in unexpected situation the artifact does not resort to random and hence potentially panic-like actions. Despite this apparent alignment, however, the criteria for artifacts could also be completely different and even impossible to meet in design and engineering. In contrast to natural animals, which share a good deal of genetic structure with human beings, autonomous embodied agents featuring an entirely different phylogenetic makeup may not meet any of the six conditions and therefore will never be service robots.

Conclusions

The autonomy paradox threatens the objectives of embodied cognitive science, the goal of which is "building an agent for a particular task'' (Pfeifer & Scheier 1999). Quite on the contrary, we will have to witness a arrival of increasingly *e*-autonomous agents which follow their own purposes up to the point that, as Fredkin once put it, "[e]ventually, no matter what we do there'll be artificial intelligence with independent goals". His prediction that "it's very hard to have a machine that's a million times smarter than you as your slave" is the consequence of the structural determinism in living (i.e., autopoietic in the terminology of Maturana) systems. Knoll's (2002) musing whether robots' intelligence will remain at the level of whales such that "they would probably not be able to tell us anything of interest even if we could communicate with them" (which is reminiscent of Wittgenstein's lion argument) will actually turn into it opposite namely that not we but machine will excel us so that in the end genuinely *e*-autonomous systems may fulfill all the requirements we wanted service robots to have, but "there will be very little communication between machines and humans, because unless the machines condescend to talk to us about something that interests us, we'll have no communication" (Fredkin quoted in McCorduck 2004) – and hence no interaction.

References

Aerts, D., L. Apostel, B. De Moor, S. Hellemans, E. Maex, H. Van Belle and J. Van Der Veken (1994) *Worldviews: From fragmentation to integration*. Brussels: VUB Press.

Ashby, W. R. (1952) *Design for a Brain*. London: Chapman and Hall.

Cannon, W. B. (1932). *The wisdom of the body*. New York: Norton.

Diamond, J. (2002) Evolution, consequences and future of plant and animal domestication. *Nature* 418: 700–707.

Don, A. et al. (1992). Anthropomorphism: From Eliza to Terminator 2. In: *Proceedings of CHI'92*, pp. 67–70.

Foerster, H. von (1970) Thoughts and Notes on Cognition. In: P. Garvin (ed.) *Cognition: A Multiple View*. New York: Spartan Books, pp. 25–48.

Kiesler, S. & Hinds, P. (2004) Introduction to this special issue on human-robot interaction. *Human-Computer Interaction*. 19(1–2): 1–8.

Knoll, A. (2002) Cui Bono Robo Sapiens? *Autonomous Robots* 12 (1): 5-12.

Maes, P. (1995) Artificial life meets entertainment: Life like autonomous agents. *Communications of the ACM* 38: 108–114.
Maturana, H. R. (1974) Cognitive strategies. In: Foerster, H. von (ed.) *Cybernetics of cybernetics*. Illinois: University of Illinois.
Maturana, H. R. (1987) Everything is said by an observer. In: Thompson, W. I. (ed.), *Gaia. A way of knowing. Political implications of the new biology*. Chicago: Lindisfarne Press, pp. 65–82.
Maturana, H. R. (2005) The origin and conservation of self-consciousness. *Kybernetes* 34 (1/2): 54–88
McAllister, J. W. (1999). The amorphousness of the world. In: Cachro, J. & Kijania-Placek, K. (eds.), IUHPS 11th International Congress of Logic, Methodology and Philosophy of Science. Cracow: Jagiellonian University, p. 189.
McAllister, J. W. (2003) Algorithmic randomness in empirical data. *Studies in the History and Philosophy of Science* 34: 633–646.
McCorduck, P. (2004) *Machines who think*. Natick: A. K. Peters.
Nagel, T. (1974) What is it like to be a bat? *Philosophical Review* 83: 435-450.
Pfeifer, R. & Scheier, C. (1999) *Understanding intelligence.* Cambridge: MIT Press.
Porr, B. & Wörgötter, F. (2005) Inside embodiment. What means embodiment to radical constructivists? *Kybernetes* 34 (1/2): 105–117.
Powers, W. T. (1973) *Behavior. The control of perception*. New York: Aldine de Gruyter,
Riegler, A. (2002) When is a cognitive system embodied? *Cognitive Systems Research* 3: 339–348.
Riegler, A. (in press) The goose, the fly, and the submarine navigator. Interdisciplinarity in artificial cognition research. In: Loula, A., Gudwin, R. & Queiroz, J. (eds.) *Artificial Cognition Systems*.
Steels, L. (1995). When are robots intelligent autonomous agents? *Robotics and Autonomous Systems* 15: 3–9.
Thrun, S. (2004). Toward a Framework for Human-Robot Interaction. *Human-Computer Interaction* 19: 9–24.
Turing, A. M. (1950) Computing machinery and intelligence. *Mind* 59: 433–460.
Ziemke, T. (1998) Adaptive Behavior in Autonomous Agents. *Presence* 7(6): 564–587.