John Benjamins Publishing Company

This is a contribution from *Pragmatics & Cognition 13:3*
© 2005. John Benjamins Publishing Company

This electronic file may not be altered in any way.
The author(s) of this article is/are permitted to use this PDF file to generate printed copies to be used by way of offprints, for their personal use only.
Permission is granted by the publishers to post this file on a closed server which is accessible to members (students and staff) only of the author's/s' institute.
For any other use of this material prior written permission should be obtained from the publishers or through the Copyright Clearance Center (for USA: www.copyright.com).
Please contact rights@benjamins.nl or consult our website: www.benjamins.com

Tables of Contents, abstracts and guidelines are available at www.benjamins.com

Robotics, philosophy and the problems of autonomy*

Willem F.G. Haselager
Nijmegen Institute for Cognition and Information (NICI)

Robotics can be seen as a cognitive technology, assisting us in understanding various aspects of autonomy. In this paper I will investigate a difference between the interpretations of autonomy that exist within robotics and philosophy. Based on a brief review of some historical developments I suggest that within robotics a technical interpretation of autonomy arose, related to the independent performance of tasks. This interpretation is far removed from philosophical analyses of autonomy focusing on the capacity to choose goals for oneself. This difference in interpretation precludes a straightforward debate between philosophers and roboticists about the autonomy of artificial and organic creatures. In order to narrow the gap I will identify a third problem of autonomy, related to the issue of what makes one's goals genuinely one's own. I will suggest that it is the body, and the ongoing attempt to maintain its stability, that makes goals belong to the system. This issue could function as a suitable focal point for a debate in which work in robotics can be related to issues in philosophy. Such a debate could contribute to a growing awareness of the way in which our bodies matter to our autonomy.

Keywords: autonomy, robotics, philosophy, embodiment

1. Introduction

Developments within Artificial Intelligence (AI) influence the way we conceive ourselves; they help to shape and change our self-image as, amongst others, intelligent and autonomous creatures. The field of robotics consists of a particularly noticeable set of tools for the study of basic features of human and animal cognition and behavior (Pfeifer 2004). In this paper I wish to approach robotics, as a cognitive technology (as described by Dascal 2004), from

Pragmatics & Cognition 13:3 (2005), **515–532**.
ISSN 0929–0907 / E-ISSN 1569–9943 © John Benjamins Publishing Company

© 2005. John Benjamins Publishing Company
All rights reserved

a philosophical perspective and examine the way it plays, as well as could play, a role in the understanding of ourselves as autonomous agents.

Over the years a growing number of claims have been made within AI regarding the existence of autonomous agents. In fact, autonomous agents appear to be a new research paradigm within AI. Candidates for autonomous agents range from software entities (such as search bots on the web) to sophisticated robots operating on Mars. At first sight, robots appear to be reasonable contenders for autonomy for several reasons. First of all, robots are embodied in the sense that their artificial bodies permit real action in the real world. In the case of evolutionary robotics, generally speaking, simulated robots operate in virtual reality, but even in these cases at least certain bodily characteristics and physical features of objects in the world are central to the simulation. Generally, within robotics the focus is on the interaction between morphology, control system and world. Thus, in robotics the emphasis is more on bodily action, whereas the traditional (purely information processing) AI approaches center more on thinking. The capacity to 'do something' makes robots relevant to the topic of autonomy. Secondly, the robots' behavior is often emergent in the sense that it is not specifically programmed for (Clark 2001), and therefore invokes characterizations as 'surprising', 'original' or even 'unwanted'. Moreover, the behavior of robots can be based on their history of interaction with the environment: they can learn (change their behavior over time as a result of the outcome of earlier actions), instead of blindly repeating fixed behavioral patterns. Thus, robots are not only doing things, but often seem to be doing them 'in their own way'. Finally, in many cases when one is observing robots it is hard to refrain from a certain amount of empathy. One immediately starts wondering about what the robot is doing or trying to achieve. Even if they fail, robots often at least seem to have struggled to achieve a goal.

Obviously, many of the words and phrases used above (like action, emergence, original, learn, and struggled to achieve a goal) are ambiguous in the sense that one could argue that these concepts only apply to robots in a metaphorical way, instead of literally, as we generally assume they do in the case of human beings and animals. For many people, robots' bodies, control systems, possibilities for learning and adaptation, and therefore ultimately their behavior is too much dependent on human programming and design in order to speak of genuine autonomy.

A complicating factor in addressing this issue of metaphorical versus literal use of the concept of autonomy is that the meaning of 'autonomy' and 'agency' is far from clear. Like so many other concepts that are central to our

© 2005. John Benjamins Publishing Company
All rights reserved

self-understanding, the notion of autonomy is difficult to define. Rather it seems to have a variety of meanings, which at best bear a family resemblance to one another, and apply to a great diversity of cases. Similarly, there does not seem to be an absolute cut-off point between autonomous and non-autonomous behavior, but rather a fuzzy border. The only thing that seems to be clear from the beginning is that too permissive and too restrictive interpretations of autonomy need to be avoided. For instance, an understanding of autonomy that would allow thermostats, or one that would only permit adult human beings to qualify, seem inadequate. We may end up with one of them, but to start the investigation on the basis of such extreme interpretations would be question begging.

Adding to the conceptual uncertainty is, as I will try to show below, the difference in interpretations of autonomy between robotics and philosophy. For some this may mean that the debate about autonomy of robots is a non-starter. I would like to suggest however, that it is precisely the interplay between empirical research in robotics and conceptual analysis in philosophy that may help to clarify the confusion. Robotics may profit from a philosophical analysis of the concept of autonomy, because this may lead to a fuller specification of the conditions that need to be fulfilled before robots can be said to have passed the benchmark. On the other hand, philosophy might gain in its attempt to clarify what is (and is not) meant by autonomy, by receiving from robotics concrete and challenging examples of behavior as test cases. Therefore, even though the confusions involved in the issue of the autonomy of robots are considerable, the debate itself can be fruitful for both philosophy and robotics. This debate, in turn, ultimately may deepen our understanding of the important role our bodies play in our autonomy.

2. Robots, autonomy and intervention

Historically, robotics grew out of the development of teleoperation (Murphy 2000), i.e., the operating of tools (often more or less like elaborate pliers) from a distance. This way, human operators could avoid having to be present in dangerous circumstances. They manipulated the operator that steered the remote, a tool that functioned for instance in an area with very high temperatures. In the late 1940s teleoperation was used in the first nuclear reactors. These teleoperators were improved by including more feedback to the human operators so they would get a better feeling of what was happening at the remote's end. Manipulating the operators to direct the remote was very tiring and time

© 2005. John Benjamins Publishing Company
All rights reserved

consuming, and soon the idea arose to have simple and repetitive operations performed automatically, without human steering. This got known as supervisory control: The operator monitors the process and gives high-level commands (e.g., 'turn') while the remote performs parts of the task on its own. All the time the human operators could take over the manipulations. This developed into part-time teleoperation and a semi-autonomous remote. It is no great exaggeration to say that the origins of the roboticist's use of the phrase 'autonomous agents' lie here. Increasing the autonomy of the remote simply means reducing the need for human supervision and intervention. Autonomy is interpreted relative to the amount of on-line (while the robot is operating) involvement of human operators.

Robots go beyond telemanipulators in that their sensorimotor capacities enable them to deal with a greater variety of circumstances and events without on-line human involvement. Robots are considered to be capable to act, in the sense that they not merely undergo events or have effects (like stones rolling down a hill). The robots' operations can be reactive, responding to what is going on (taxes and tropisms), but also proactive, in pursuit of the goals that are active within them, thereby becoming less environmentally driven. Such proactive robots can, as Nolfi and Floreano (2002: 31) put it, be "let free to act", in the sense that they can choose how to achieve these goals.

Fanklin and Graesser (1996) provide a review of the various interpretations of autonomous agents that currently circulate within AI. Rather than repeating that review,[1] I will offer a definition that, I think, captures the general intent: *Autonomous agents operate under all reasonable conditions without recourse to an outside designer, operator or controller while handling unpredictable events in an environment or niche.*

The 'without recourse to an outside designer' refers to recourse *during* the act (i.e., on-line), but not to recourse to a designer preceding the behavior (i.e., programming). Here it is usually pointed out that there is a continuum between complete dependence and complete independence (e.g., Maes 1995: 108). The 'under all reasonable conditions' is added to indicate that there are limits to the robot's functioning ('reasonable'), while at the same time signifying that the robot should not be too dependent on favorable circumstances ('all'). The clause about unpredictable events in an environment is intended to rule out pre-configured systems operating blindly in a completely predetermined environment. Within robotics, then, the increase in autonomy of a system is related to the reduction of on-line supervision and intervention of the operator, programmer or designer in relation to the robot's operations in a changing environment.

© 2005. John Benjamins Publishing Company
All rights reserved

3. Agents and goals

In the philosophical literature, however, one finds rather more emphasis on the reasons *why* one is acting (i.e., the goals one has chosen to pursue) than on *how* the goals are achieved. Auto-nomos, being or setting a law to oneself, indicates the importance of self-regulation or self-government. Autonomy is deeply connected to the capacity to act on one's own behalf and make one's own *choices*, instead of following goals set by other agents. The importance of being able to select one's own goals is also part and parcel of the common sense interpretation of autonomy.

Although it is impossible to do justice here to all the historical complexities involved, one can trace an opposition between causation through choice versus physical causation back to Plato and Aristotle. In the *Phaedo* (98c–99b), Socrates argues that an explanation of his sitting or lying down purely in terms of his bones, sinews and joints, i.e., in terms of physical or necessary (or, what Aristotle would call efficient) causation, is missing the real cause, namely the reason for his sitting, which is based on Socrates' aim, the result of his choice of what is best to do. Aristotle, in his discussion of the four kinds of causes, emphasized the importance of final causation: "there is the goal or end in view, which animates all the other determinant factors as the best they can attain to; for the attainment of that 'for the sake of which' anything exists or is done is its final and best possible achievement" (*Physics* II, iii; 195a 24–26). For Aristotle, choices are the result of deliberate desires to do something; it is through deliberation that we consider how to put our objectives into practice, and our choices reveal who we are (Hutchinson 1995: 208–210). The essential characteristic of voluntary behavior is that the origin of movement is within the agent's psyche (Juarrero 1999: 16–19).

From this perspective, then, the conception of autonomy within robotics is not very satisfactory. Robots may be operating independently — even 'freely' choosing how to act in order to achieve goals — but the goals they are trying to achieve are still set by human programmers. Although the reduction of on-line involvement of human operators is a considerable achievement for robotics, one could doubt that it is sufficient for the autonomy of the robots involved, because of the enormous amount of human off-line involvement (i.e., the programming and designing of the robot) in advance of the robot's functioning, particularly in relation to which goals the robots should pursue. The contrast between philosophy and robotics regarding this issue is perhaps best illustrated by considering the following set of examples (it is not uncommon to find such

© 2005. John Benjamins Publishing Company
All rights reserved

cases in philosophical papers on autonomy (see, for instance, Mele 1995; Dennett 2003: 281–285; Mele 2004).

A case of *full* (or harmonious) *autonomy* would be the following: I weigh the pros and cons of drinking beer for a long time, decide that home-made is fine, so I brew my own beer and drink it. This situation is different from a case of *strong will*, where I'd like to drink beer, but since I want to loose weight, I take water instead. This, in turn, differs from the perhaps more familiar case of *weakness of will*; I want to be healthy and think that drinking beer is detrimental to my health, but I buy beer in a shop and drink it all the same. Then there are more extreme cases, such as *delusions*, where I may think that drinking beer is the only way to thwart a plot of aliens to take over the world, so I go to a bar and order a beer. Finally, there is the philosophically popular example of being *brainwashed*. In such a case, external agents make me think that I want to drink beer because it is good for me (or to stop the aliens, or whatever), so I decide to sneak into a beer-factory at night and drink the whole lot.

In these examples, one goes from autonomy of will (selecting a goal) and action (performing a specific behavior) to autonomy of action (weakness of will), to no autonomy at all. Accordingly, the amount of responsibility for my actions decreases (even in the case of delusions I may be considered to bear some responsibility, as I could have taken therapy or medicine, while this option is not available in the case of brainwashing). On the basis of these examples, a philosopher might argue that robots are in a situation comparable to that of people who are brainwashed and that therefore robots are *not even close* to being candidates for any serious degree of autonomy. Robots are not autonomous because they themselves don't choose their own goals and they do not even know it is us that set their goals for them. In relation to the 'true' (philosophical) meaning of autonomy, robots are on the other end of the spectrum.

4. Freely choosing goals?

However, that such a conclusion would be a bit premature becomes clear if one considers some developments in the 17th century philosophical debate about autonomy and the ability to choose one's goals. During more or less one century, the clear distinction between choice and necessity started to become quite vague and the notion of final cause fell into disrepute. In relation to the latter, Juarrero (1999: 2) says that purposive, goal-seeking, final causation "no longer even qualified as causal; philosophy restricted its understanding of causality

© 2005. John Benjamins Publishing Company
All rights reserved

to efficient cause". Efficient causality refers to the push and pull kind of impact of forces on inert matter and it has dominated explanatory practices since the 17th century.

In relation to the former, Vesey (1987: 17) indicates that Descartes introduced a concept of the will that went beyond the classical Greek interpretation of will. According to Vesey, the concept of will was interpreted as 'adopting a favorable attitude to some specific object'. For Descartes, however, 'will' referred to a separate faculty with the power to cause voluntary movements, so that "from the simple fact that we have the desire to take a walk, it follows that our legs move and that we walk (1649: 340). He equated the will with freedom of choice and said: "The will simply consists in our ability to do or not do something" (1641: 40).

However, as soon as volitions are considered to be the causes of actions, Vesey (1987: 20) notes: "it is hard not to allow that volitions are caused by, say, motives, that motives are caused by character, and that character is caused by heredity and environment". Descartes insisted that the will was free so that the chain of causes would end immediately ("I cannot complain that the will or freedom of choice which I received from God is not sufficiently extensive or perfect, since I know by experience that it is not restricted in any way." (1641: 39). But Hume went beyond Descartes in claiming that the acts of the will are caused by motives according to the bonds of necessity:

> We may imagine we feel a liberty within ourselves; but a spectator can commonly infer our actions from our motives and character; and even where he cannot, he concludes in general, that he might, were he perfectly acquainted every circumstance of our situation and temper, and the most secret springs of our complexion and disposition" (Hume 1739: 408–409; III, ii).

Not much later, Hartley (1749: 12; Prop. IX) defended a mechanicism, neurophysiologically grounded in his theory of the 'vibrations of medullary particles' according to which

> each action results from the previous circumstances of body and mind, in the same manner, and with the same certainty, as other effects do from their mechanical causes; so that a person cannot do indifferently either of the actions A, and its contrary a, while the previous circumstances are the same; but is under an absolute necessity of doing one of them, and that only (Hartley 1749: 84; Conclusion).

Motives are 'the mechanical causes of actions' (1749: 86; Conclusion). So, within the century separating Hartley from Descartes, the contrast between events

© 2005. John Benjamins Publishing Company
All rights reserved

caused by choices versus events caused by necessity had all but disappeared. Although philosophy started out with a conception of agent causation and the capacity to freely choose goals, it ended up in the 18th century with a strikingly mechanistic view. Vesey claims that the natural outcome of this development was formulated most radically in the 20th century by the psychologist Skinner (1953: 447–448) who wrote the infamous *Beyond Freedom and Dignity* (1971) and claimed that the ultimate causes of behavior lie outside the agent and that 'man is not free'.

5. Goals: Having them vs. choosing them

At this point, it would not be unreasonable for roboticists to shrug their shoulders and claim that it is not reasonable to expect a solution to the problem of freedom of will through the development of robots. After all, the issue of freedom of will has turned into one of the major issues in philosophy and it would be far fetched to expect current research in robotics to solve a metaphysical problem that philosophy itself is still trying to come to terms with. As Strawson (1998) says about the problem of free will: "New generations (…) will doubtless continue to launch themselves onto the old metaphysical roundabout". Similarly, as noted earlier, the technical problem of greater independence does not deal with the philosophically interesting aspects of autonomy. These two different versions of the problem of autonomy provide little ground for a debate between robotics and philosophy. In fact, one could even argue that robotics fails as a cognitive technology in relation to autonomy because the gap between what is done and what (philosophically speaking) should be done is too large. Robots, from this perspective, do not provide significant means for the production of useful knowledge concerning autonomy.

However, it is possible to distinguish one more version of the problem of autonomy. This version concerns the question, how and when the goals of creatures genuinely become *theirs*. My goals, at least my basic ones, really belong to me. But when, and on what basis, could we say that robots are pursuing goals of their *own*? This issue of intrinsic *ownership* has to be separated from the harder problem of the freedom of will, i.e., whether we are free to choose what we want to do, and how that would be compatible with physical determinism (and it also has to be distinguished from the technical problem of autonomy within robotics, as described above).

© 2005. John Benjamins Publishing Company
All rights reserved

As said, I think it is unrealistic to expect solutions in relation to the problem of free will from robotics. The question concerning the ownership of goals seems to me to lend itself more to input from robotics because, as I will venture below, it raises questions about the *integration* between body, control system, and the aims of the actions undertaken by the system. This is something that can be, and has been, studied both conceptually and empirically.

Even if I don't freely choose my goals, the goals I pursue are *mine*. My goals are important and intrinsically connected to me. Certain goals I do not pursue because others impose them upon me or ask me to achieve them, but because they *matter* to *me*. What makes goals belong to a system? How are they grounded in the system? In the following I will offer a suggestion, that may help to sketch a possible path towards giving an answer to these questions, and I will attempt to show how this can be related to research in robotics.

Fundamentally, what makes my goals mine, is that I myself am at stake in relation to my success or failure in achieving them.[2] That is, goals belong to a system when they arise out of the ongoing attempt, sustained by both the body and the control system, to maintain homeostasis. To a significant extent, it is the body, and its ongoing attempt to maintain its stability, that provides the founding of goals within the system. Autonomy is grounded in the formation of action patterns that result in the self-maintenance of the embodied system and it develops during the embodied interaction of a system with its environment. There are two aspects involved in this suggestion that bear some further investigation in relation to robotics. First of all, there is the issue of the integration between the control system and the body. Secondly, the notion of homeostasis deserves a closer look.

6. The integration between body and control system

The biologist von Uexküll (1864–1944) stressed the importance of the integration of all of the organism's components into one purposeful whole. We have to see, he claimed

> in animals not only the mechanical structure, but also the operator, who is built into their organs as we are into our bodies. We no longer regard animals as mere machines, but as subjects whose essential activity consists of perceiving and acting (Uexküll 1957: 6).

Furthermore, he stressed that machines act according to plans of their human designers, but that living organisms are acting plans (Uexküll 1928: 301; see

© 2005. John Benjamins Publishing Company
All rights reserved

also Ziemke and Sharkey 2001: 708). It is this 'building the operator into the body' that provides, I believe, a profound but legitimate challenge to robotics in relation to the problem of the ownership of goals.

As Chiel and Beer (1997) have pointed out, the brain and the body have developed in constant conjunction during their evolutionary and lifetime interaction with the environment. The search, then, is for an approach that allows for a tight coupling between bodies and control-systems both phylogenetically and ontogenetically. Against this background, the field of evolutionary robotics, with its aim to drive the programmer and designer 'out of the robot' as much as possible, is a very interesting recent development (see, e.g., Sims 1994a 1994b; Nolfi and Floreano 2000).

6.1 Evolutionary robotics

Evolutionary robotics has been defined as "the attempt to develop robots and their sensorimotor control systems through an automatic design process involving artificial evolution" (Nolfi 1998: 167). Artificial evolution involves the use of genetic algorithms. The 'genotypes' of robots are represented as bits that can code their morphological features as well as the characteristics (such as weights and connections of a neural network) of their control systems. A fitness formula determines candidates for reproduction by measuring the success of the robots on a specific task. The genotypes of the selected robots are then subjected to crossover with other genotypes and further random mutation, giving rise to a new generation of robots. According to Nolfi (1998: 167–168), the organization of the evolving systems is the result of a self-organizing process, and their behavior emerges out of the interactions with their environment. Therefore, evolutionary robotics is relevant to the topic of autonomy since there is less need for the programmer and/or designer to 'pull the strings' and shackle the autonomy of the evolving creatures, because the development of robots is left to the dynamics of (artificial) evolution.

Artificial evolution is far from straightforward, however, and usually requires an extensive amount of preparation before the evolutionary process can take off. As Nolfi (1998: 179) points out:

> In principle (…) the role of the designer may be limited to the specification of a selection criterion. However, (…) in real experiments the role of the designer is much greater than that: In most of the cases the genotype-to-phenotype mapping is designed by the experimenter; several parameters (e.g., the number of individuals in the population, the mutation and crossover rate, the length of

© 2005. John Benjamins Publishing Company
All rights reserved

> the lifetime of each individual, etc.) are determined by the experimenter; and in some cases the architecture of the controller is also handcrafted. In theory, all these parameters may be subjected to the evolutionary process; however, in practice they are not.

Moreover, it is not unusual to find that the designer not only pre-arranged the evolutionary process, but also interfered directly with its course in order to solve thorny issues such as local minima and the bootstrap problem.[3] In relation to problems such as these, Nolfi (1997) reports, for instance, having to change the fitness formula by adding elements that cause reward for behavior that in itself is not very meaningful, and increasing the number of encounters with relevant stimuli. Often, then, "some additional intervention is needed to canalize the evolutionary process into the right direction" (Nolfi 1997: 196), keeping the designers well in control of their robots, even if the strings to which the robots are tied may be less visible. A second difficulty concerns the fact that, in practice, what evolves is often just the control system and not the body (the morphology of the robot). One of the most used robots is the khepera, a small pre-made robot, with specific sensori-motor capacities that can be controlled through neural networks. In the context of artificial evolution, most often khepera simulators are used instead of the real robots (for reasons of time and costs). The neural networks operate simulated bodies, which are similar in certain respects to the khepera, in a virtual world. After the neural networks have gone through a number of changes during the artificial evolutionary process they can be downloaded into real khepera in a process that could be described as a simple form of 'brain transplantation'. It is important to realize that during the evolutionary process, the khepera itself (both the real one and its simulated counterpart) does not undergo any changes at all. There is no equivalent for this in real evolution. If the integration between body and control system is important for autonomy, it is legitimate to doubt approaches that focus on evolving a neural network in relation to a fixed and pre-designed robot body.

However, there is a growing amount of research on the co-evolution of body and control system. Sims (1994a, 1994b) and Harvey, Husbands, and Cliff (1994) provide early examples, and have worked with simulations of robots and environments. More recently, Pollack, Lipson, Hornby, and Funes (2001) have evolved real robots by means of a 3 dimensional printer that uses thermoplastics to build bodies in a flexible way. Here, then, we have body-control system co-evolution and a greater (though by no means complete) emphasis on real (vs. simulated) robots. However, Pollack et al. (2001: 11) note that the types of robots that could be built this way are fairly simple, and conclude:

© 2005. John Benjamins Publishing Company
All rights reserved

> The limitations of the work are clearly apparent: these machines do not yet have sensors, and are not really interacting with their environments. Feedback from how robots perform in the real world is not automatically fed-back into the simulations, but require humans to refine the simulations and constraints on design. Finally, there is the question of how complex a simulated system can be, before the errors generated by transfer to reality are overwhelming (Pollack et al. 2001: 14).

Let me summarize. The 'ownership version' of the problem of autonomy leads to the consideration of how goals become grounded in systems. A suggestion I have pursued is that this grounding is based in part on the integration of all bodily components into one purposeful, homeostasis oriented, whole. From this perspective, robotics is interesting because of the developments within evolutionary robotics that aim to model the phylogenetic co-development of the body-control system. Moreover, these developments are accompanied by growing possibilities for translating the simulations into hardware versions, thereby further emphasizing the importance of the interaction between real bodies and real environment. At the same time, however, one has to acknowledge the considerable technical difficulties encountered by recent projects in co-evolutionary robotics. Although it is far too early to draw any clear conclusions, this aspect of the ownership interpretation of autonomy at least creates the possibility for a fruitful debate between philosophers and roboticists about co-evolutionary developed robots.

7. Homeostasis: How bodies matter

The notion of homeostasis refers to the regulation of the internal environment of open systems to remain within a region of stability. The concept was introduced by Claude Bernard (1813–1878) and the term by the biologist Walter Cannon in 1932 (meaning: same — steady, i.e., to remain the same). For instance, when glucose concentrations in blood are too high, receptors in the pancreas start a process that results in the release of the hormone insulin, stimulating the conversion of glucose into glycogen that can be stored in the liver, thereby decreasing the glucose concentration. This is simply put, of course, and other examples include oxygen, temperature, water, and urea. This capacity for self-regulation in the service of self-maintenance is characteristic of living organisms. Importantly, homeostasis involves more than just keeping a variable constant through the use of feedback, as in the case of a thermostat regulating temperature, in that the homeostatic system necessarily *depends,* for its own

© 2005. John Benjamins Publishing Company
All rights reserved

existence, on the self-regulation. A malfunctioning or incorrectly set thermostat need not suffer from the negative consequences it produces, but a truly homeostatic system always will.

Clearly, the notion of homeostasis plays a fundamental role within robotics. Specifically, the relation between homeostasis and self-maintenance is well studied. A simple but perhaps illustrative example concerns the regulation of the amount of energy that a robot has at its disposal. The robot regularly checks its energy level, and takes the necessary actions when the supply runs dangerously low, e.g., by returning on time to its battery reload station. If this kind of self-checking and self-maintaining mechanism does not operate properly, the robot necessarily will suffer from the consequences, as it will simply stop functioning.

However, I would like to suggest that the type of homeostasis that is at issue here is merely functional, but *not genuinely embodied*. To clarify this distinction, let me start by pointing out a well-known difference between robots and living organisms. One can turn robots off for an indefinite amount of time and start them later without any principled problems. A similar procedure is, as we all know, impossible in the case of living creatures. Once 'turned off', they stay off. This is, at least in part, due to the fact that the bodies of current robots are fundamentally different in kind compared to the type of bodies belonging to living organisms. Organic matter decays when not part of a functioning whole, whereas the plastics and metals of robots suffer no such fate.

I think that Maturana and Varela's (1987) concept of *autopoiesis* is particularly relevant to deepen our understanding of this difference. Autopoiesis refers to the self-generating and self-maintaining capacity of the basic building blocks of organic bodies: cells. As Ziemke and Sharkey (2001: 733) say, living organisms consist of autopoietic unities, self-producing and self-maintaining systems. An autopoietic system is a homeostatic machine, and the fundamental variable it aims to maintain constant is *its own organization*. This makes an autopoietic system different from homeostatic machines whose bodies can continue to exist even if they stop operating. The self-organizing capacity of living bodies is based on the autopoietic quality of their basic elements. Such a quality is missing in current robot bodies. Perhaps another way to point to the same difference is to note that an autopoietic system aimed at homeostasis needs to interact continually, for as long as it exists, with its environment. Its basic goals, the ones that really matter to it, *enforce* this continuous interaction (on pains of annihilation). For currently existing robots, the type of homeostasis they are aimed at demands no such thing.

© 2005. John Benjamins Publishing Company
All rights reserved

One may think that, again, it would not be unreasonable for roboticists to shrug their shoulders and say that they can hardly be expected to work with organic material and living creatures. There is, after all, a difference between robotics and biology. However, my plea for genuinely embodied homeostasis should *not* be taken as a request to build robots out of living cells, for the notion of autopoiesis does not reflect some intrinsic quality of a specific kind of matter but rather indicates a characteristic of the *organization* of matter. As Maturana and Varela (1987: 51) say: "the phenomena they generate in functioning as autopoietic unities depend on their organization and the way this organization comes about, and not on the physical nature of their components" (see also Ziemke and Sharkey 2001: 732).

Given that it is the organization of the components and not their material constitution that matters, the question is open whether autopoiesis could be realized in artificial matter. All things considered, I do not think that autopoiesis provides a *principled* obstacle for robotics. The argument does indicate a further constraint on robotics, however.[4] Currently robots are constructed mainly out of metals and plastics. A question that needs to be pursued is whether these types of materials allow for a genuinely autopoietic organization. In a way, this brings back Aristotle's ideas about the relationship between matter and form. The form can actualize the potentialities of the matter, but the potentialities have to be there: you cannot build a boat out of sand. A more thorough investigation of the relationship between homeostasis and autopoiesis may lead to the development of what perhaps might be called a *mild* functionalism that pays attention to material aspects of the body to a higher degree than currently is customary. Again, a fruitful debate between philosophers and roboticists concerning this point seems possible.

8. Conclusion

Robotics constitutes a valuable cognitive technology because it helps in the understanding of our selves as autonomous agents, also when it is, at times, premature in laying certain claims. Debates about the differences between machines and organisms can further sharpen our knowledge about what actually constitutes autonomy. Analyzing the debate between roboticists and philosophers, I have tried to indicate that they have different conceptions of autonomy, emphasizing, respectively, the capacity for independent (unsupervised) action versus the freedom to choose goals. I have also pointed out some

© 2005. John Benjamins Publishing Company
All rights reserved

possible historical reasons for this difference. In the course of this paper, I have distinguished three different problems of autonomy. The first one concerns the technical problem of how one selects the right type of behavior to achieve a certain goal. The second problem concerns the hard problem of freedom of will, whether (and if so, how) it is possible to freely choose one's own goals. Finally, there is the issue of how and when goals genuinely belong to the creature itself, instead of merely being imposed upon, or installed within, it. I have suggested that the first problem lacks philosophical import, while the second is out of reach for empirical approaches such as robotics. The third one, however, is relevant for robotics and of considerable philosophical interest as well. Regarding this third problem, one possibility worthy of further investigation is that the capacity to have goals of one's own arises out of the continuous integration of control system and body, resulting in actions aiming at homeostasis. A further understanding of this capacity requires considering the co-evolution of body and control system as well as the specific 'potentialities' of organic matter, i.e., autopoiesis. In relation to both aspects, an exchange between empirical research and conceptual analysis seems possible and potentially fruitful. A collaborative investigation of philosophy and robotics of autonomy might lead to a further strengthening of the growing acknowledgement within cognitive science of the importance of our material constitution for cognition. In relation to our continuing attempt to understand who we are, bodies may matter even more than we currently think.

Notes

* I would like to thank Susan van den Braak and especially Iris van Rooij for detailed and helpful comments on earlier versions of this paper.

1. Five definitions that are particularly relevant (though not all equally illuminating) here are the following.

(1) Franklin and Graesser (1996: 5): "An autonomous agent is a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future". (2) Brustoloni (1991,in Franklin 1995: 265): "Autonomous agents are systems capable of autonomous, purposeful action in the real world". (3) Wooldridge and Jennings (1995: 2): "autonomy: agents operate without the direct intervention of humans or others, and have some kind of control over their actions and internal state". (4) Maes (1995: 108): "Autonomous agents are 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". (5) Murphy (2000: 4): "'Functions autonomously' indicates that the robot

© 2005. John Benjamins Publishing Company
All rights reserved

can operate, self-contained, under all reasonable conditions without requiring recourse to a human operator. Autonomy means that a robot can adapt to changes in its environment or itself and continue to reach its goal".

2. Of course, there are many 'high-level' goals that are not directly or profoundly related to 'being at stake' (e.g., when I set myself the goal to redecorate the living room). I suggest a similar type of relation between such high-level goals and the more basic ones as between the social (e.g., shame or pride) and more basic (fear or happiness) emotions. That is, in order for the high-level goals to be genuinely mine, the existence of more basic goals that are grounded in my body-control system integration aimed at homeostasis is required.

3. Local minima arise when after a certain amount of progress in relation to the performance of a task, new generations stop improving and the evolving robots get stuck in a sub-optimal performance. The bootstrap problem involves how to get beyond the starting point when the individual robots of the initial generation are unable to differentiate themselves in relation to the task because they all score zero, so that there is no way to select the 'best' or least worst individuals.

4. Pfeifer (2004: 120) provides another reason to emphasize the importance of the materials used for robots: "Most robot arms available today work with rigid materials and electrical motors. Natural arms, by contrast, are built of muscles, tendons, ligaments, and bones, materials that are non-rigid to varying degrees. All these materials have their own intrinsic properties like mass, stiffness, elasticity, viscosity, temporal characteristics, damping, and contraction ratio to mention but a few. These properties are all exploited in interesting ways in natural systems". He argues that robotics similarly should take advantage of the intrinsic properties of matter, in order to simplify the tasks for control systems. Although I am in complete agreement with his argument, my point differs from his in the sense that 'natural matter' not just simplifies the control problem, but also plays a role in grounding the very goals the control system is trying to achieve.

References

Aristotle. *Physics*. Loeb edition. Cambridge, MA: Harvard University Press.

Brustoloni, J.C. 1991. "Autonomous agents: Characterization and requirements". *Carnegie Mellon Technical Report CMU-CS–91–204.* Pittsburgh: Carnegie Mellon University

Chiel, H.J. and Beer, R.D. 1997. "The brain has a body: Adaptive behavior emerges from interactions of nervous system, body and environment". *Trends in Neurociences* 20(12): 553–557.

Clark, A. 2001. *Mindware: An Introduction to the Philosophy of Cognitive Science.* Oxford: Oxford University Press.

Dascal, M. 2004. "Language as a cognitive technology". In B. Gorayska and J.L. Mey (eds), *Cognition and Technology: Co-existence, Convergence and Co-Evolution.* Amsterdam: John Benjamins, 37–62.

Dennett, D. 2003. *Freedom Evolves*. New York: Viking.

© 2005. John Benjamins Publishing Company
All rights reserved

Descartes, R. 1984 [1641]. *Meditations on First Philosophy*. In J. Cottingham, R. Stoothoff, and D. Murdoch (eds), *The Philosophical Writings of Descartes, Vol. 2*. Cambridge: Cambridge University Press, 3–62.

Descartes, R. 1967 [1649]. *Passions of the Soul*. In E.S. Haldane and G.R.T. Ross (eds), *The Philosophical Works of Descartes, Vol. 1*. Cambridge: Cambridge University Press, 330–427.

Franklin, S. 1995. *Artificial Minds*. Cambridge, MA: The MIT Press.

Franklin, S and Graesser, A. 1996. "Is it an agent, or just a program? A taxonomy for autonomous agents". In *Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages*. Berlin: Springer, 21–35.

Hartley, D. 1970 [1749]. *Observations on Man*. In R. Brown (ed), *Between Hume and Mill: An Anthology of British Philosophy 1749–1843*. New York: The Modern Library, 5–92.

Harvey, I., Husbands, P., and Cliff, D. 1994. "Seeing the light: Artificial evolution, real vision". In D. Cliff, P. Husbands, J.A. Meyer, and S. Wilson (eds), *From Animals to Animats III*. Cambridge, MA: The MIT Press.

Hume, D. 1978 [1739]. *A Treatise of Human Nature*. Oxford: Clarendon Press.

Hutchinson, D.S. 1995. "Ethics". In J. Barnes (ed), *The Cambridge Companion to Aristotle*. Cambridge: Cambridge University Press, 195–232.

Juarrero, A. 1999. *Dynamics in Action: Intentional Behavior as a Complex System*. Cambridge, MA: The MIT Press.

Maes, P. 1995. "Artificial life meets entertainment: Life like autonomous agents". *Communications of the ACM* 38(11): 108–114.

Maturana, H.R. and Varela, F.J. 1987. *The Tree of Knowledge: The Biological Roots of Human Understanding*. Boston: Shambhala.

Mele, A. 1995. *Autonomous Agents*. Oxford: Oxford University Press.

Mele, A. 2004. "Dennett on freedom". *Electronic publication*. Accessed at 29–12–2004. http://gfp.typepad.com/online_papers/files/Al.doc.

Murphy. R.R. 2000. *Introduction to AI Robotics*. Cambridge, MA: The MIT Press.

Nolfi, S. 1997. "Evolving non-trivial behaviors on real robots: A garbage collecting robot". *Robotics and Autonomous Systems* 22: 187–198,

Nolfi, S. 1998. "Evolutionary robotics: Exploiting the full power of self-organization". *Connection Science* 10(3–4): 167–184.

Nolfi, S. and Floreano, D. 2000. *Evolutionary Robotics: The Biology, Intelligence and Technology of Self-Organizing Machines*. Cambridge, MA: The MIT Press.

Nolfi, S. and Floreano, D. 2002. "Synthesis of autonomous robots through evolution". *Trends in Cognitive Sciences* 8(1): 31–37.

Pfeifer, R. 2004. "Robots as cognitive tools". In B. Gorayska and J.L. Mey (eds), *Cognition and Technology: Co-Existence, Convergence and Co-Evolution*. Amsterdam: John Benjamins, 109–126.

Plato. *Phaedo*. Loeb edition. Cambridge, MA: Harvard University Press.

Pollack, J.B., Lipson, H., Hornby, G.S., and Funes, P. 2001. "Three generations of automatically designed robots". *Artificial Life* 7(3): 215–223.

Sims, K. 1994a. "Evolving virtual creatures". *Siggraph '94 Proceedings*, 15–22.

Sims, K. 1994b. "Evolving 3D morphology and behavior by competition". In R.A. Brooks and P. Maes (eds), *Artificial Life IV*. Cambridge, MA: The MIT Press, 28–39.

© 2005. John Benjamins Publishing Company
All rights reserved

Skinner, B. 1953. *Science and Human Behavior.* New York: The Free Press.
Skinner, B. 1971. *Beyond Freedom and Dignity.* New York: Bantam Books.
Strawson, G. 1998. "Free will". In E. Craig (ed), *Routledge Encyclopedia of Philosophy*. London: Routledge. Retrieved January 24 2005, from http://www.rep.routledge.com/article/V014SECT5
Uexküll, J. von 1957. "A stroll through the worlds of animals and men: A picture book of invisible worlds". S*emiotica* 89(4): 319–391.
Vesey, G. 1987. "Plato's two kinds of causes". In A. Flew and G. Vesey (eds), *Agency and Necessity*. Oxford: Basil Blackwell.
Wooldridge, M. and Jennings, N.R. 1995. "Agent Theories, architectures, and languages: A survey". In M. Wooldridge and R. Jennings (eds), *Intelligent Agents*. Berlin: Springer, 1–22.
Ziemke, T. and Sharkey, N.E. 2001. "A stroll through the worlds of robots and animals: Applying Jakob van Uexküll's theory of meaning to adaptive robots and artificial life". *Semiotica* 134(1/4): 701–746.

Author's address

Willem F.G. Haselager
Artificial Intelligence / Cognitive Science
Nijmegen Institute for Cognition and Information (NICI)
Radboud University
B.02.35, Spinozagebouw, Montessorilaan 3
6525 HR Nijmegen
The Netherlands

w.haselager@nici.ru.nl
http://www.nici.ru.nl/~haselag
http://www.notedpages.blogspot.com

About the author

Willem Haselager is Assistant Professor of Artificial Intelligence / Cognitive Science at the Nijmegen Institute for Cognition and Information (NICI), Radboud University. He is a regular visiting professor at the Philosophy Department of the Universidade Estadual Paulista (UNESP), in Marília, SP, Brazil. He holds master degrees in philosophy and psychology and a Ph.D. in theoretical psychology. He is particularly interested in the integration of empirical work (i.e., psychological experiments, computational modeling, and robotics) with philosophical issues regarding knowledge and intelligent behavior. He analyzed the debate between proponents of classical cognitive science and connectionism on the nature of representation, in relation to the inability of computational models to deal with the frame problem (interpreted as related to abduction and common sense knowledge and reasoning) and examined the consequences of that debate for the status of folk psychology. More recently he has extended his research by investigating the embodied embeddedness of cognition (EEC) in relation to dynamical systems theory (DST) and the theory of self-organization. Of late, he has started to explore issues in conscious will and autonomy.

© 2005. John Benjamins Publishing Company
All rights reserved