Sign Systems Studies 30.1, 2002

On the epigenesis of meaning in robots and organisms: Could a humanoid robot develop a human(oid) Umwelt?

Tom Ziemke

Department of Computer Science, University of Skövde
Box 408, 54128 Skövde, Sweden
e-mail: tom@ida.his.se

Abstract. This paper discusses recent research on humanoid robots and thought experiments addressing the question to what degree such robots could be expected to develop human-like cognition, if rather than being pre-programmed they were made to learn from the interaction with their physical and social environment like human infants. A question of particular interest, from both a semiotic and a cognitive scientific perspective, is whether or not such robots could develop an experiential Umwelt, i.e. could the sign processes they are involved in become intrinsically meaningful to themselves? Arguments for and against the possibility of phenomenal artificial minds of different forms are discussed, and it is concluded that humanoid robotics still has to be considered "weak" rather than "strong AI", i.e. it deals with models of mind rather than actual minds.

Even readers with no interest whatsoever in the scientific and philosophical study of artificial intelligence (AI) might have noticed the following: Back in 1968, in Stanley Kubrick's movie *2001 — A space odyssey*, it was the spaceship's board *computer* HAL whose intelligence exceeded by far that of his human collaborators. Now that we have actually reached the year 2001 the appearance of AI in popular culture has taken a significantly different shape. In Steven Spielberg's recent movie *A. I.* (based on a treatment of Stanley Kubrick, who died before he could produce the movie himself), it is the humanoid *robot* David, looking very much like any ordinary little boy, who exhibits not only human-level intelligence, but also develops human feelings (or at least convincingly appears to do so).

Obviously there is a huge gap here between the science fiction and the actual science facts. Neither computers like HAL nor robots like David have been built or could be built within the foreseeable future. Nevertheless it might be worth pointing out a couple of parallels to actual AI research. From its inception in the mid-1950s AI, as well as the rest of the cognitive sciences, was dominated by the so-called *computer metaphor for mind*, which stated that cognition is computation and the relation between mind and brain/body the same as between computer software and hardware. Accordingly, an understanding of mind was sought not at the level of the body, which was considered just an implementation — which happens to be carbon-based in the case of humans, but could as well be silicon-based in the case of computers — but at the level of implementation-independent representations and algorithms. That means, given the right program, i.e. the program used by the human mind, a computer like HAL could indeed have a human mind.

This view has been strongly contradicted by, among others, Searle (1980) who in his famous *Chinese Room Argument* compared a computer's processing of internal symbols/representations to a non-Chinese-speaking man's processing of Chinese symbols according to formal rules without grasping any of the semantics. In both cases, Searle argued, the symbol processing might very well be meaningful to observers, but it cannot possibly be or become intrinsically meaningful to the processor itself. Hence, computers might very well be powerful tools in the study of cognition, a position Searle referred to as *weak AI*, but they could not be actual minds themselves, a position he referred to as *strong AI*. Searle (1980) did, however, not conclude that strong AI in general, i.e. the building of artificial minds, was impossible, but only that computer programs are the wrong approach due the fact that they lack a number of "causal powers", including *perception*, *action* and *learning*.

Since the late 1980s increasingly many cognitive scientists, to some degree following Searle's ideas, have emphasized the importance of "embodiment" and "situatedness", i.e. interaction of cognitive agents with their environments (e.g. Varela *et al.* 1991; Clark 1997;

Clancey 1997; Pfeifer, Scheier 1999). AI has been one of the driving forces in this development, shifting much interest from computers to robots or so-called *autonomous agents*, and from the study of internal knowledge representation to sensorimotor processes and the way they shape cognition. One of the insights gained (or regained) was that the mind is in fact not largely independent of the body, but in fact strongly determined by it. Not surprisingly, Uexküll's concepts of *Umwelt* and *Merkwelt* have been adopted by a number of AI researchers and cognitive scientists (e.g., Brooks 1986, 1991; Prem 1996, 1997, 1998; Clark 1997; Sharkey, Ziemke 1998; Ziemke, Sharkey 2001; Ziemke 2001). Brooks (1991), for example, writes: "as von Uexküll and others have pointed out, each animal species, and clearly each robot species with its own distinctly nonhuman sensor suites, will have its own different Merkwelt". For AI research striving to model human intelligence this has radical consequences: Clearly, if cognition is dependent on body and sensorimotor capacities, then the only way to achieve or study human-level or human-like intelligence in artefacts is to equip them with human-like bodies and sensorimotor capacities, i.e. to build *humanoid robots*.

There are by now a number of projects which have taken this approach, such as Brooks' well-known *Cog* project (Brooks *et al.* 1998) or Kozima's *Infanoid* project (e.g. Kozima, Yano 2001). Both Cog and the Infanoid are upper-torso humanoids, i.e. roughly human-size robotic torsos equipped with stereo-vision heads, arms and hands with degrees of freedom roughly similar to those of human bodies. However, obviously this only solves part of the problem. Even if a (human-like) body nowadays by many is considered a *necessary* condition for a (human-like) mind, it could hardly be a *sufficient* one. The remaining question is, roughly speaking, how to get a mind "into" the body. Both of the above projects aim to let their robots undergo some kind of *artificial ontogenesis* in physical and social interaction with their environment. Both also particularly emphasize the interaction with human caregivers, based on theories of social learning in infants (e.g., Vygotsky 1978; Tomasello 1999). That means, Cog and Infanoid are supposed to acquire or develop sensorimotor and cognitive capacities, and ultimately a mind, in some kind of long-term interaction similar to the ontogenesis of human children (note, however, that it is only the software, not the hardware/body, which develops). Taking this approach to the extreme, one might argue like

Zlatev (2001: 155) that such "robotogenesis could possibly recapitulate [human] ontogenesis, leading to the emergence of intentionality, consciousness and meaning" in robots.

The question whether or not a (humanoid) robot could indeed develop/have a (human-like) mind, including a (human-like) phenomenal Umwelt, has recently occupied a number of researchers in cognitive science and semiotics (e.g., Emmeche 2001; Nöth 2001; Sharkey, Ziemke 1998; Ziemke, Sharkey 2001; Ziemke 2001; Zlatev 2001). The question what exactly the semiotic status of such a robot would be apparently has no simple answer. Traditionally, semiosis has often been considered to necessarily involve living organisms. Morris (1946), for example, defined semiosis as "a sign-process, that is, a process in which something is a sign to some organism". Similarly, Jakob von Uexküll considered signs to be "of prime importance in all aspects of life processes" (T. von Uexküll 1992), and made a clear distinction between organisms, which as *autonomous subjects* respond to signs according to their own *specific energy*, and inorganic mechanisms which are *heteronomous* (cf. Nöth 2001; Ziemke, Sharkey 2001).

Nowadays, the distinction between organisms and mechanisms seems less clear. Computers are commonly considered to be at least involved in semiotic processes. Sebeok, for example, writes (in personal communication cited by T. von Uexküll 1982) that "the criterial feature of living entities, and of machines programmed by humans, is semiosis". Andersen *et al*. (1997) have argued in detail that computers/programs, when it comes to semiosis, fall somewhere in between humans and conventional mechanisms, but that they ultimately derive their semiotic "capacities" from the interpretation of their designers and users. The major difference, they argued, was that living systems are autopoietic, i.e. self-creating and -maintaining, whereas machines are not (cf. Nöth 2001; Ziemke, Sharkey 2001). Hence, their "tentative conclusion" was that "the difference between human and machine semiosis may not reside in the particular nature of any of them. Rather, it may consist in the condition that *machine semiosis presupposes human semiosis and the genesis of the former can be explained by the latter*" (Andersen *et al*. 1997: 569, emphasis added). Similarly, Nöth concluded his discussion of whether or not robots have an *Umwelt* as follows:

> Needless to say, a machine, in spite of a certain autonomy in its agency, can never be said to have its ultimate goal within itself. The objectives of a machine have always been established from outside, namely by the engineer who designed it and the user who switches it on and off. Thus, the robot's ultimate framework of reference, its final causality, is elsewhere, and thus the resulting semiotic process is alloreferential. (Nöth 2001: 696–697)

However, many would argue that in the case of robots which self-organize and develop in long-term interaction with their environment, independent of their human designers, it is simply not the case that the genesis of robosemiosis can be (fully) explained with reference to human semiosis. The "problem" that makes it difficult, at least at a first glance, to make a sharp distinction between living organisms and today's adaptive robots (also commonly referred to as *artificial life*), is that the latter nowadays have a number of the qualities/properties of the former. Ziemke and Sharkey (2001), for example, discussed in detail that three properties which Jakob von Uexküll (1928, 1982) considered unique for organisms (adaptation/growth, use of signs, centrifugal construction) can to some degree also be found in today's robots. Similarly, Nöth (2001: 695–696) identified "four reasons why robots interact in the same way with their environment as organisms do" which "support the argument that not only organisms, but also robots have an Umwelt in [von] Uexküll's sense": (a) both robots and organisms have an Umwelt (or in fact Merkwelt) in the sense that, limited by available senses/sensors, they can only sense part of their physical environment; (b) both process environmental stimuli selectively; (c) both can have "internal representations of their Umwelt"; (d) both are equipped with perceptual organs/modules and effector organs/modules.

Given these similarities between robots and organisms, arguments for the possibility of robot minds cannot easily be dismissed. Zlatev, for example, sees "no good reason to assume that intentionality is an exclusively biological property […] and thus a robot with bodily structures, interaction patterns and development similar to those of human beings would constitute a system possibly capable of meaning" (Zlatev 2001: 155). In more detail, Zlatev's elaborate proposal for the development of a robot mind[1] is based on the following cornerstones:

[1] It should be noted that this proposal is fairly similar to the ideas underlying both Infanoid and Cog project.

> (*) sociocultural situatedness: the ability to engage in acts of communication and participate in social practices and 'language games' within a community;
>
> (*) naturalistic embodiment: the possession of bodily structures giving adequate causal support for the above, e.g. organs of perception and motor activity, systems of motivation, memory and learning; […]
>
> (*) epigenetic development: the development of physical, social, linguistic skills along a progression of levels so that level n+1 competence results from level n competence coupled with the physical and social environment. (Zlatev 2001: 161)

In the case of a robot that actually fulfilled all of the above criteria it might indeed be difficult to justify why exactly it should not be considered to have a human-like mind and Umwelt. It might very well pass what Harnad (1989, 1990) called the *Total Turing Test,* i.e. its behavior, including both symbolic capacities (as tested in the original, purely language-based Turing test) as well robotic, i.e. sensorimotor, capacities, might become indistinguishable from that of a human. Nevertheless, according to Nöth (2001), it is just a man-made machine, lacking own goals and thus only capable of "alloreferential" semiotic processes (cf. above quote). Nöth's argument, as well as our own arguments coming to similar conclusions (Sharkey, Ziemke 1998; Ziemke, Sharkey 2001; Ziemke 2001), might seem counterintuitive, as can be demonstrated with the following thought experiment (in fact an extension of Zlatev's (2001) thought experiment). Let us assume you buy some future version of Cog or Infanoid, now equipped with legs, etc., so it does actually look like a child (perhaps even as much as Spielberg's fictitious humanoid David). Let us further assume that the robot learns, e.g., through language games (cf. Zlatev 2001) to refer to your family, your dog and objects in your house by their proper names. Could we really say, as Nöth (cf. above quote) seems to argue, that its language use and all other semiotic processes are *alloreferential,* i.e. the words have no intrinsic meaning to the robot itself, but they are only meaningful to you and your family? What if the robot, unknown to you, played with the neighbor's children and learned new words and phrases from them, or possibly even went to school? Finally, what if eventually it could pass the Total Turing Test? Is there really any good reason to assume that such a robot should not be able to develop own intentionality and intrinsic meaning?

Well, there are in fact a couple of good reasons, and here are some of them. Firstly, although the above robot seems to possess at least some form of the "causal powers" that Searle (1980) pointed out as missing in computer programs (cf. above), i.e. perception, action and learning, the Chinese Room Argument (CRA) still applies to it. As Cog and Infanoid (cf. above), the robot consists of hardware and software. It has a physical body and a computer program, or perhaps a number of programs, controlling it. Each of these programs is of exactly the type Searle (1980) argued to be incapable of intentionality due to their computational nature,[2] and their embedding in a robot (the so-called *robot reply*) is exactly what he rejected as making no difference whatsoever. It should, however, be pointed out that, of course, not everybody agrees with Searle in this point (see, e.g., Harnad 1990; Zlatev 2001).

Secondly, despite a certain convergence of science fiction and philosophical thought experiments, it should be pointed out that the above is indeed just a thought experiment. Its technical feasibility does in fact seem more than questionable. The idea that a humanoid robot could develop a human mind and Umwelt, just because its body is to *some* degree human-like and thus might be able to, e.g., receive similar visual input and have similar possibilities of, e.g., manually grasping objects, seems to reduce the body to some kind of input-output interface to the world. Robot bodies are, however, in many ways extremely different from living bodies, in particular human bodies, and thus unlikely candidates for supporting the same kind of phenomenal mind/Umwelt. In particular, robot bodies (hardware) and control systems (software) are not at all integrated the way living bodies are. Robot bodies do, for example, not grow. Furthermore, Ziemke and Sharkey (2001) argued in detail that robots lack *endosemiosis* and therefore also lack what T. von Uexküll *et al.* (1993) referred to as the *neural counterbody*, formed and updated in our brain as a result of the continual information flow of proprioceptive signs from the muscles, joints and other parts of our limbs, and thus giving rise to the experience of the living body as the center of our subjective reality. That means, even if you believe that such a humanoid was capable of exhibiting human-like behavior and having *a* phenomenal

[2] As pointed out by Searle (1990), this includes connectionist/neural networks.

Umwelt, exactly what reasons are there to believe that the Umwelt *would* be human-like?[3]

Does this mean that artificial minds (in the strong sense) are impossible? Of course it does not. Our conclusion from the first above argument is just like that of Searle (1980), that AI might very well be possible, but not with central cognitive processes implemented as computer programs, i.e. purely formally defined systems. The conclusion from the second above argument is that, taking embodiment seriously, and taking the bodily differences seriously, (a) humanoids are due to the lack of integration between body and software unlikely to be able to exhibit human-like behavior, and (b) even if they could, they would still be unlikely to do so with a human-like mind.

As discussed in detail elsewhere (Sharkey, Ziemke 1998, 2001; Ziemke, Sharkey 2001; Ziemke 1999, 2001), we believe that the key to understanding mind is to understand the *autonomous* and *autopoietic*, i.e. self-creating and -maintaining, nature of living systems (Maturana, Varela 1980). Autopoietic systems have a natural (rather than a metaphysical) kind of intentionality or aboutness in the sense that they are autonomous unities concerned with assimilation/dissimilation of material from/into their environment for the purpose of self-maintenance and survival. Living systems are also far more integrated than the above humanoids in the sense that their ontogenesis does in fact start from a single cell from which they grow in a centrifugal fashion (Uexküll 1982; cf. Ziemke, Sharkey 2001; Ziemke 2001). Hence, a more natural route towards artificial minds would be the attempt to create artificial autopoietic systems (cf. also Boden 1999). This would be very unlikely to result in systems even remotely similar to humans, but it would avoid the somewhat dualist/functionalist approach of building a hardware body and then trying to make it develop a software mind.

In sum, it has been argued here that robots, as long as they are allopoietic machines consisting of “dead” hardware bodies and computational control programs, will not be able to develop intrinsic meaning or autonomy by means of some kind of artificial ontogenesis as envisioned by Zlatev (2001). The sign processes embedding living systems into their environment, on the other hand, as well as their

[3] Elsewhere we have discussed in detail the relation to the case of Clever Hans (Sharkey, Ziemke 2001).

ontogenetic development, are intrinsically meaningful to themselves due to their autopoietic, self-creating and -maintaining nature.[4]

References

Andersen, Peter B.; Hasle, Per; Brandt, Per A. 1997. Machine semiosis. In: Posner, Roland; Robering, Klaus; Sebeok, Thomas A. (eds.), *Semiotics: A Handbook on the Sign-Theoretic Foundations of Nature and Culture*. Berlin: Walter de Gruyter, 548–571.

Boden, Margaret 1999. Is metabolism necessary? *British Journal of the Philosophy of Science* 50(2): 231–248.

Brooks, Rodney A. 1986. Achieving artificial intelligence through building robots. *Technical Report Memo 899*. Cambridge: MIT AI Lab.

— 1991. Intelligence without representation. *Artificial Intelligence* 47: 139–159.

Brooks, Rodney A.; Breazeal, Cynthia; Marjanovi, Matthew; Scasselati, Brian; Williamson, Matthew 1998. The Cog Project: Building a Humanoid Robot. In: Nehaniv, Chrystopher L. (ed.), *Computation for Metaphors, Analogy, and Agents*. New York: Springer, 52–87.

Clancey, William J. 1997. *Situated Cognition: On Human Knowledge and Computer Representations*. New York: Cambridge University Press.

Clark, Andy 1997. *Being There: Putting Brain, Body and World Together Again*. Cambridge: MIT Press.

Emmeche, Claus 2001. Does a robot have an Umwelt? *Semiotica* 134(1/4): 653–693.

Harnad, Stevan 1989. Minds, machines and Searle. *Journal of Experimental and Theoretical Artificial Intelligence* 1: 5–25.

— 1990. The symbol grounding problem. *Physica D*, 42: 335–346.

Kozima, Hideki; Yano, Hiroyuki 2001. A robot that learns to communicate with human caregivers. In: *First International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*. (Lund University Cognitive Studies vol. 85.) Lund, 47–52.

Morris, Charles W. 1946. *Signs, Language, and Behavior*. Englewood Cliffs: Prentice Hall.

Nöth, Winfried 2001. Semiosis and the Umwelt of a robot. *Semiotica* 134(1/4): 695–699.

Pfeifer, Rolf; Scheier, Christian 1999. *Understanding Intelligence*. Cambridge: MIT Press.

Prem, Erich 1996. *Motivation, Emotion and the Role of Functional Circuits in Autonomous Agent Design Methodology*. Technical Report 96–04. Vienna: Austrian Research Institute for Artificial Intelligence.

[4] *Acknowledgments*. The author would like to thank Noel Sharkey, Jordan Zlatev, Jessica Lindblom and Ron Chrisley for discussions that have contributed much to this paper. This work was supported by a grant (1507/97) from the Knowledge Foundation, Stockholm.

— 1997. Epistemic autonomy in models of living systems. In: *Proceedings of the Fourth European Conference on Artificial Life*. Cambridge: MIT Press, 2–9.

— 1998. Semiosis in embodied autonomous systems. In: *Proceedings of the IEEE International Symposium on Intelligent Control*. Piscataway: IEEE, 724–729.

Searle, John 1980. Minds, brains and programs. *Behavioral and Brain Sciences* 3: 417–457.

— 1990. Is the brain's mind a computer program? *Scientific American* January: 20–25.

Sharkey, Noel E.; Ziemke, Tom 1998. A consideration of the biological and psychological foundations of autonomous robotics. *Connection Science* 10(3/4): 361–391.

Sharkey, Noel E.; Ziemke, Tom 2001. Mechanistic versus phenomenal embodiment: Can robot embodiment lead to strong AI? *Cognitive Systems Research* 2(4): 251–262.

Tomasello, Michael 1999. *The Cultural Origin of Human Cognition*. Cambridge: Harvard University Press.

Uexküll, Jakob von 1928. *Theoretische Biologie*. Berlin: Springer.

— 1982. The theory of meaning. *Semiotica* 42(1): 25–82.

Uexküll, Thure von 1982. Introduction: Meaning and science in Jakob von Uexküll's concept of biology. *Semiotica* 42(1): 1–24.

— 1992. Introduction: The sign theory of Jakob von Uexküll. *Semiotica* 89(4): 279–315.

Uexküll, Thure von; Geigges, Werner; Herrmann, Jörg M. 1993. Endosemiosis. *Semiotica* 96(1/2): 5–51.

Uexküll, Thure von; Geigges, Werner, and Herrmann, Jörg M. 1997. Endosemiose. In: Posner, Roland; Robering, Klaus; Sebeok, Thomas A. (eds.), *Semiotik: Ein Handbuch zu den zeichentheoretischen Grundlagen von Natur und Kultur*. Berlin: Walter de Gruyter, 464–487.

Varela, Francisco J.; Thompson, Evan; Rosch, Eleanor 1991. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge: MIT Press.

Vygotsky, Lev S. 1978. *Mind in Society: The Development of Higher Psychological Processes*. Cambridge: Harvard University Press.

Ziemke, Tom 1999. Rethinking Grounding. In: Riegler, Alex; Peschl, Markus; Stein, Astrid von (eds.), *Understanding Representation in the Cognitive Sciences*. New York: Plenum Press, 177–190.

— 2001. The construction of 'reality' in the robot: Constructivist perspectives on situated artificial intelligence and adaptive robotics. *Foundations of Science* 6(1): 163–233.

Ziemke, Tom; Sharkey, Noel E. 2001. A stroll through the worlds of robots and animals: Applying Jakob von Uexküll's theory of meaning to adaptive robots and artificial life. *Semiotica* 134(1/4): 701–746.

Zlatev, Jordan 2001. The epigenesis of meaning in human beings, and possibly in robots. *Minds and Machines* 11: 155–195.

О эпигенезе у роботов и организмов: может ли у человекоподобного робота развиться человекоподобный Umwelt?

Статья рассматривает новейшие исследования, связанные с человекоподобными роботами, и мыслительные эксперименты, занимающиеся вопросом, до какой степени у подобных роботов может развиться человекопободное сознание, если вместо запрограммированности начинать их обучать как детей, посредством общения со своей физической и социальной средой. Особенно интересен вопрос (как в семиотической так и когнитивно-научной перспективе), может ли таким образом у роботов выработаться основанный на опыте Umwelt, т.е. могут ли знаковые процессы, в которых они участвуют, стать внутренне значимыми для них самих? Рассматриваются аргументы как за, так и против возможности разных форм искусственного интеллекта и делается вывод, что область человекоподобных роботов нужно считать скорее “слабым” чем “сильным искусственным интеллектом”.

Tähenduse epigeneesist robotitel ja organismidel: kas inimsarnasel robotil võiks areneda inim(sarnane)-omailm?

Käesolev artikkel käsitleb uuemaid uurimusi inimsarnaste robotite vallas ning mõtte-eksperimente, mis tegelevad küsimusega, mil määral seesugustel roboteil võiks eeldatavasti areneda inimsarnane teadvus, kui ette programmeerituse asemel panna nad õppima — suhtlemise kaudu oma füüsilise ja sotsiaalse keskkonnaga, nagu inimlapsed. Iseäranis huvipakkuv küsimus (nii semiootilisest kui ka kognitiivteaduslikust perspektiivist) on, kas seesugustel roboteil võiks areneda kogemuslik omailm, s.t kas märgiprotsessid, milles nad osalevad, võiksid saada neile enestele sisemiselt tähenduslikuks? Käsitletakse nii poolt- kui vastuargumente tehisvaimu erinevate vormide võimalikkuse suhtes ning järeldatakse, et inimsarnaste robotite valdkonda tuleks pigem pidada “nõrgaks” kui “tugevaks tehisintellektiks”.