Agents vs. Rossum's Robots: Towards Intelligent Living Machines[*]

Matěj Hoffmann[1], Cyril Brom[2]

Charles University, Faculty of Mathematics and Physics
Malostranské nám. 2/25, Prague, Czech Republic
[1]matej.hoffmann@seznam.cz
[2]brom@ksvi.mff.cuni.cz

Abstract. In this paper, we rethink the problem of intelligent living machines in the context of original Robots from the play R. U. R. (Rossum's Universal Robots) by Karel Čapek. We analyze Robots in terms of recent artificial intelligence paradigms – cognitivism, emergentism and enactive viewpoint –, and highlight conceptual positives and mistakes of Robot inventors. Finally, we draw a line between Robots and current research in AI and discuss which novel research directions could be fruitful and which could end in an impasse. The aim is not to criticize any research efforts, rather, to isolate some problems with the hopes of furthering debate.

1 Introduction

Some forty years ago, the world of artificial intelligence (AI) was represented among others by systems playing chess or diagnosing illness, systems based on manipulation with formal symbols. Later, connectionism has entered the scene and our most basic assumptions about intelligent machines had to be changed. Not long ago, a small but loudly speaking research community has introduced enactive viewpoint and this new concept has been forcing us to revise once again our notion of how cognition and intelligence could be understood. At the same time, the concept of artificial intelligent agents has been born of a symbiosis between AI and software engineering.

Do we know yet whether some of the past approaches in AI are obsolete? Which ones and in what context? Why?

A generation before the work of McCulloch and Pitts [12], which is generally recognized as first in AI, Karel Čapek had introduced his famous play R. U. R. (Rossum's Universal Robots, [6], 1921), which gave origin to the word 'robot'. The aim of the playwright was to address the problem of mankind [5, Chap. 18]. Čapek wanted to depict the threat of dehumanization and showed the heroic and unique character of human beings.

We think that Čapek's play has also a lot to say to problems the present AI faces, although the author's attention was originally focused rather on human, not science issues. In this paper, we rethink the paradigms of modern AI in the context of original

[*] Both authors of the paper are students.

Robots[1]. Above all, cognitivism, emergentism and enactive viewpoint will be analyzed and present AI systems and artificial agents will be brought face to face with Robots. Our motivation is simple. We want to clarify the issues on real, intelligent living machines and on understanding of cognition in the hopes of furthering debate on this topic. We think that a powerful way to achieve this objective can be through the study of relations between approaches of current AI researchers and successes and mistakes of Robot inventors.

In the following tour, we make four stops. First, we recall R. U. R. Second, we analyze Robots from the point of view of aforementioned paradigms. We anticipate that the first generation of Robots can be considered as created in an emergentist manner and the second generation to resemble the enactive viewpoint. Third, we link Robots with intelligent agents and AI systems. And finally, we propose three directions of possible future research in the field of intelligent machines.

2 R. U. R.

In this section, we briefly look through the developments of the play, concentrating on the evolution of Robots and motivations of main characters. There are two generations of Robots in the play. The following characters are crucial for the history of Robots:

- OLD ROSSUM[2]: father of Robots; the inventor of living matter
- YOUNG ROSSUM: his son; the inventor of the first generation of Robots
- DIRECTOR DOMIN: director of R. U. R. factory
- HELENA GLORY: his wife, ‘mother’ of the second generation of Robots
- DR. GALL: the head of the research department of the factory

The first two characters, old and young Rossum, are already dead in the course of the play. We know about them from the story by director Domin.

Old Rossum was representing *materialism* of the 19th century. He discovered a *living matter*, a mysterious substance, which became the basis for all creatures that were made. This matter was made chemically. As director Domin puts it:

> Old Rossum wanted to scientifically dethrone God… He wanted nothing but to furnish evidence that there was no need for any God. He thus set his heart upon making a man to a hair identical to us. [6, p.122][3]

However, Rossum was making his man for ten years but he lived for three days only. Rossum made another two beings during the rest of life, but with the same flaw.

Young Rossum had a different ambition. He wanted to make “living and intelligent working machines.”[4] He adopted an effective *engineering approach* and hence director Domin could say that [6, p. 124]: “he represented the age of production

[1] Capital ‘R’ is used when talking about the Robots from the play – in accordance with the text.

[2] The name resembles the Czech word ‘rozum’, meaning intelligence, reason.

[3] All quotes from [6] and [9] are translated from Czech by the authors.

[4] This corresponds to the Slavic origin of the word Robot. ‘Robota’ meaning serfdom, obligatory work in Czech.

after the age of science." Young Rossum realized that human anatomy could be simplified a lot for this purpose. Emotions, spiritual needs, reproduction could have been left out easily. In the words of director Domin:

> Young Rossum invented a worker with minimum needs. He had to simplify him. He rejected everything that did not serve directly the purpose of work. He thus in fact discarded man and made a Robot….They are mechanically superior to us, they have an incredible rational intelligence, but lack a soul. [6, p. 124-125]

The details of the construction are not mentioned but the chemical living matter was combined in an engineering way to produce only the inevitable parts of a worker. Such simplification resulted in a success and gave rise to the first generation of Robots. They were then produced on a large scale and distributed as workers to the whole world.

Helena Glory originally came to the factory to fight for the Robots' rights. Then, she became the wife of director Domin and stayed in the factory. Helena had a very humane attitude. Her empathy with the Robots and her wish that they should have a soul made her convince Dr. Gall to change them. It was this view that resulted in the evolution of true autonomous living beings, as Robots Primus and Helena demonstrated at the very end of the play. It was also this second generation of Robots, which lead a successful revolution against people on Earth.

3 Čapek's Robots and present-day AI paradigms

In this section, we characterize three co-existing and competing paradigms in present-day AI with the help of Robots. The purpose is to reveal some interesting contexts of the paradigms and to clear the ground for further discussion in Section 4.

3.1 Cognitivism

The first paradigm is called *cognitivism* or *functionalism* (or computer functionalism) as well as *strong AI*. The thesis of strong AI may be formulated as follows:

> The nature of mind is algorithmic and it is not significant in which media are the algorithms (programs) implemented." [9, p. 33]

The guiding metaphor is a digital computer, which performs computations on symbols representing features of the world. Hence, the keywords of this paradigm are *algorithmic nature*, *symbolic processing* and *representations*.

We suggest that none of the Robots were built in a cognitivist manner. It is not mentioned in the play whether Robots perform any symbolic manipulation and represent external world (this is not to be wondered at with respect to the time of origin of the play). Nevertheless, it is clearly evident that it is the living matter that makes the Robots live. This is the fundament and hence the Robots are not independent of the media. That means, they would not work without the living matter—they are not programs, which could be loaded to another media.

3.2 Emergentism and the first generation of Robots

The second paradigm is *emergentism* or *connectionism*. It rests on the idea that many complex phenomena, such as autonomous intelligent behaviour of a robot, are emergent properties of parallel events in a network of simple active elements with links between them [9, p. 38]. The most popular example of active parts is neurons and connectionism is usually 'connected' with these. Nevertheless, these active parts can be in principle anything. It is possible to say that these are pieces (or 'cells') of artificial living matter in the case of R. U. R. In the following, we will talk about emergentism since this term is not associated with neurons.[5] We anticipate that the first generation of Robots in R. U. R. can be considered as made in the emergentist manner.

Production of Robots. The development method utilized by young Rossum was functional decomposition. Tissues and organs were made from the living matter and assembled into a Robot. After that, emergence came into play. As Domin explained:

> ...after the living matter is prepared, individual organs, bones, nerves and veins are manufactured. Then they (Robots) are assembled like cars. After that, they are put in a store, where they get into existence. They internally adhere somehow. Many things even newly grow in them. You see, we have to leave some time for natural development." [6, p. 129-130]

It is possible to say that the artificial living matter and the organs *emerged* into a Robot after some time. Robots were then taught to speak, write and calculate and finally they were prepared for work. [6, p. 130]

Zombie machines. In the course of prelude of the play, it was clearly revealed that Robots of young Rossum were only machines. They were designed just to *pretend* human-like behaviour. Robot Sulla working as the director's secretary was so convincing in imitating a human counterpart that Helena thought she was a girl like herself. However, she lacked any internal drives; she did not know what is joy, sorrow or fear of death. The human characters in the play denied the possibility of this Robot having any inner experience or any feelings. As we hear from the directors of the factory, they "never think of anything new" and:

> They are just Robots. Without own will. Without passion. Without history. Without soul. [6, p.136]

And without love and defiance. These Robots also did not feel pain. When looking at them from outside, they could be considered as falling into the category of goal-based agents as described in Russell and Norvig [16, p. 42-44].[6] The reason for this is

[5] Havel [9] accents that in living organisms the situation is not that simple; there is not just one 'lower' level and emerging properties at the 'upper' level, but rather a hierarchy of levels. Moreover, the levels may not be distinct and it may also be difficult to say which level is above which.

[6] They are rather goal-based describable, not goal-driven. *Goal-driven* stands for agents and robots that manipulate with symbols representing their goals, while *goal-based describable* stands for creatures that could be described as such only from outside – the internal mechanism is not known.

simple: Robots felt no satisfaction when accomplishing a goal; they followed commands without any notion of their utility, without any motivation. They were just checking off goals on a pregiven list (see Fig. 1). If Robots received any feedback from environment, they could not have understood it. Despite the initial period of learning and emergence, they remained just "zombie-machines".

World model. We suggest that, since Robots had neither emotions nor drives, their view of world was not changing. Their *internal model of the world*, if they had any, could at best accumulate bundles of facts without any meaning for the Robots themselves.

After considering these facts, we conclude that these Robots are not intelligent living systems, which live their own life. They are cheap and reliable working mechanisms in the human world, nothing more.

3.3 Enaction and the second generation of Robots

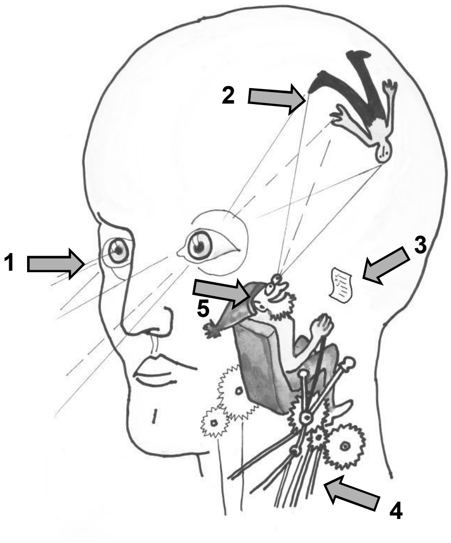
We approach the enactive viewpoint from the natural humane perspective of Helena Glory. She intuitively saw very well what the Robots were missing.

> ...if only they could be given a little bit of love – [6, p. 132]
>
> Why don't you create a soul to them? [6, p. 137]

The initiative of Helena finally resulted in the creation of truly living beings; she became the mother of the second generation of Robots.

Enactive viewpoint has been introduced in the early ninetieths as a counterstroke to classical cognitive and emergentist approaches in AI, which could not cope with some issues on inner experiencing of organisms, feelings or real understanding [20]. In the

Fig. 1: Typical notion of architecture of a cognitive being.
Sensors (1) receive percepts (2), that are further processed by a central procedure and eventually memorized (5). The procedure is accomplishing some goals from the pregiven list (3), while commanding effectors (4). Effectors change the underlying physical structure of the pregiven body. Notice, that this scheme fits both symbolist and connectionist artificial intelligent agents.
However, the depicted creature is not embodied, even though it is not bodiless. Do you think that the homunculus inside the head feels anything but feedback from the joystick and backache?



cognitivist approach, the robot constructors impose their model of the world on the agents and robots (the model is typically simplified for the robot's particular tasks). Thus, the robot has a representation of the world that he has not acquired himself. In emergentism, the conditions are also prepared from outside. Then, typically under human supervision, by telling the robot what is right and wrong (from the perspective of the teacher!), a cognitive system emerges.

Thus, according to these views, the agent (or organism) is parachuted into a pregiven world, which he tries (with a certain help from the constructor) to internally represent. Varela *et al.* [20] offer a different view. The agent is viewed not as an input/output machine but rather as a network consisting of multiple levels of interconnected, sensorimotor subnetworks that are *embodied* in the environment. Cognition is then viewed as *embodied action*.

> By using the term *embodied* we mean to highlight two points: first, that cognition depends upon the kinds of experience that come from having a body with various sensorimotor capacities, and second, that these individual sensorimotor capacities are themselves embedded in a more encompassing biological, psychological, and cultural context. By using the term action we mean to emphasize once again that sensory and motor processes, perception and action, are fundamentally inseparable in lived cognition. [20, p. 172-173]

The history of sensorimotor patterns (action and perception) gives rise to cognitive structures. There is a continual *enactment*, shaping of the world: through perceiving and acting in the world the world changes the agent and hence the agent's world is changed. The agent's cognition is a result of *coupling with the environment that brings forth a world*. The term "world" stands rather for natural or internal world of the organism, not the "all what is outside". These "own worlds" are relative to each species and individual being. An individual inherits a long history of structural coupling of his ancestors.

Could we consider the second generation of Robots to be enactive beings?

Second generation. The birth of the second Robot generation was started by Dr. Gall, who added nerves for pain to the Robots. The reason was to make the Robots more careful with their body, so that they did not damage themselves. We think that this can be interpreted as the first step towards embodied cognition. The Robots were able to 'enact' what is dangerous in the world.

Later in the play, Helena convinced Dr. Gall to give soul to the Robots. He explained what he had done afterwards:

> I have changed the nature of the Robots, and their production. Only some physical conditions, you see? Above all – above all – their – irritability. [6, p. 181]

The Robots have gained a feedback from the "what is outside". Not from a supervisor. They began to truly feel, what the realm around them is like. They started to act and perceive in their own world. As one of the Robots put it [6, p. 203]: "We were machines, Sir; but from horror and pain we have become – ...souls."

In our opinion the second Robot generation can be interpreted as created under enactive robotics. The Robots changed and learned through their living in the world. In the term of Varela [20]: the history of *structural coupling* with the environment has begun. From that time on, the autonomy of Robots could only grow (see Fig. 2).

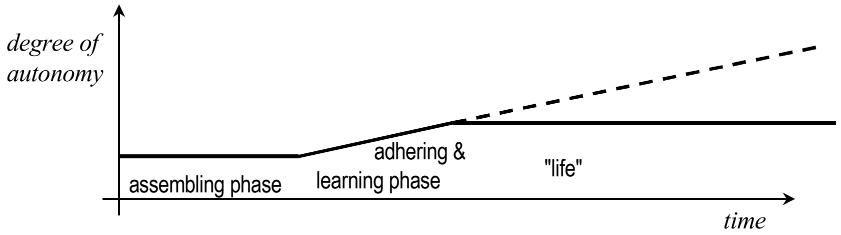


Fig. 2: "Degree" of autonomy and enactment of Robots. It is symbolised here that the first generation of Robots was autonomous only to a small degree, although in the course of learning the autonomy might have increased. Dash line depicts the continuous evolution of the second generation of Robots. Notice, that the term "autonomy" is used in the sense of Ziemke [25].

4 Connections between Robots and present-day AI research

We have so far given a 'guiding tour' around recent paradigms in artificial intelligence exploiting Robots as a demonstration example. Could we also discover some analogies between Robots and particular results in current AI research? In what follows, we briefly visit three examples of such interesting links. The aim is not to criticize any research efforts, rather, to set out ground for furthering debate on the topic of novel research directions. Notice that the term *intelligent agent* will be used in the following text in the sense of Wooldridge [23].

Ontologies and semantic web. Old Rossum was making his creature for ten years, but it died after only three days. He tried to build a man from scratch, he "hooked" it to a pregiven external world, which the creature could not know about! He wanted to dethrone God, but he forgot that God might have exploited evolution in order to *ground* organisms in their environment. We think that this act of old Rossum resemble current attempts to create agents that are aimed at understanding natural world of humans (although these agents typically lack a body) – read newspapers or web articles, for example. The original cognitive approach how to attain this goal was to create (and eventually standardize) huge symbolic ontologies of human knowledge and exploit some kind of inference. Project Cyc [10] is an example of this approach. This multi-million dollar attempt to build a system that would exhibit real common sense was generally viewed as failed as early as ten years ago [24]. Nevertheless, such applications are becoming increasingly popular again in the domain of semantic web. In particular, world-net activity and building of synonym sets (synsets) knowledge bases [8] resemble building of omniscient artificial golems.

It should be noted that the point of discussion is neither to discard semantic web efforts nor ontology research. Rather, our message is that it is necessary to think carefully about what is an attainable goal and what is an impossible mission. Surely, in the very near future, there will exist some agents that would act intelligently in a "narrow-domain", solving for example physical therapist searching problem in the web area [3]. But agents built on pure symbolic methodology, that would exhibit common sense in its entirety, will hardly ever exist. The problem is that these agents neither act in natural world nor explore it. They are even not coupled with natural environment through bodies. Instead, they do some deterministic computation within a pile of abstract symbols prearranged in some relations. Even if they learn (using for example so-called *chunking* from Soar architecture of Newell e*t al.* [14]), they obtain nothing new about the world, but facts derived from pregiven description. There is neither experience nor feeling here.

Embodied agents. Unlike pure symbolic agents, first generation Robots could be considered *embodied*. They act directly in the environment using their bodies and have immediate feedback on their own sensation. Nevertheless, their bodies do not evolve, except for a short period in a store where they "internally adhered somehow". The Robot employs his own body, but again, it is only a pregiven component designed by engineers (see Fig. 1). Surprisingly, the same problem holds for robots of Brooks [4], which are often viewed as typical representatives of enactive beasts. As Ziemke [25] noted, Brooks and his followers have neglected a simple fact: every natural organism is deeply historically rooted in its environment through the species history and the development of the living body of the particular individual. Representational-less robots with the pregiven body could be best described as puppets controlled by "environmental puppeteers". Moreover, robots and agents controlled by neural networks have often the same problem. The networks are adaptive but they are just control structures, not bodies themselves.

We suggest that from the point of view of the challenge of truly intelligent machines there is a need for research on evolution of underlying robots' physical structures. First steps towards this have already been taken (e.g. [11]). Another promising direction might be building creatures directly from a "living matter". One example of such an effort is a computer agent controlled directly by biological neural network cultivating on a Petri dish [1]. However, the results here are very premature.

Artificial evolution. Enaction is a history of *structural coupling* of agent and its environment that brings forth a world [20]. The evolution of Robots, that starts with Primus and Helena is not only an evolution of a new generation. It is also an evolution of a new world (*i.e.* internal world of Robots). Love, for example, becomes a new quality of it. Accordingly, we suggest that open-ended artificial evolutions like these of Tierra system [15] have to include evolution of both agents and environment. Not only does an agent have to receive the feedback from its environment, but also vice-versa. If the environment is invariable, artificial evolution reaches its limit soon. Hence, we think that the research in environments for multi-agent systems [21] might be fruitful, as well as the focus on agent-environment interaction and dynamical approach to it, e.g. [2].

Emotions. They should not be neglected when constructing intelligent living machines. It is evident that humans do not behave optimally according to a utility function. Rather, they behave *suboptimally*. Several researchers suggest that emotions might be fundamental in deciding which goal to pursue, *e.g.* Minsky in [13, Chap. 16] and in the new book he is preparing. Emotions could also significantly increase *believability* of agents that "only" imitate humans, see for example [7, Chap. 6, 19].

5 Conclusion: Steps towards living machines?

In this paper, we have gone through the play of R. U. R. (Rossum's Universal Robots) by Karel Čapek. The reason is simple: we think that the play can say a lot to present-day science, because it addresses many philosophical, ethical, theoretical and even practical issues.

Particularly, we have discussed the notion of embodiment of modern artificial intelligent agents and robots, as well as cognitive, connectionist and enactive AI paradigms in the context of original Robots. We have shown some directions in the current research that, at least as we think, might be fruitful with respect to the objective of building real intelligent living machines. On the other hand, we have isolated some research trends, whose goals together with methods would be better to rethink over carefully.

We conclude that there are three possible ways before us, researchers. Firstly, we could follow an engineering approach represented by young Rossum and start to build machines. That means Wooldridge-like intelligent autonomous agents that solve tasks in narrow domains in order to make life for us, humans, easier. That is an honest mission. All machines of piece have been built with this objective. Nevertheless, they are just machines. Only such expectation can be considered rational, that would anticipate a software system to exhibit such amount of common sense like, for example, a crane does.

Secondly, we could try to create real intelligence. No matter whether an agent should think as we do or in a different manner, a novel approach must be used. Should the creature demonstrate real understanding of its environment, it must be grounded in it; it must really act in it using its own body, not a bundle of bolts, sheets and wires pre-arranged by a creator. To paraphrase Havel [9, p. 53]: To achieve this goal, researchers should stop building machines that pretend they are humans or organisms. They must start to build humans and organisms. Only these could be able to "enact" their own world. However... could such beasts be useful?

A third approach has not been discussed in this paper. Nevertheless, it is of significant importance. It is the so-called *synergistic approach* noted by Shaw [18, p. 41] as having its roots at California Institute of Technology in early seventies. They claimed that computers should not have been treated as autonomous devices, rather they should have been coupled with humans to outperform capabilities of both. Efforts like these of Warvick might be considered as steps in this direction. Could we create cyborgs?

We think that all ways are feasible. Nevertheless, from the research point of view, the commitment to one of them must be done unambiguously. It is not possible to go right and left at the same time.

Acknowledgements. This research is partially supported by the Program "Information Society" under the project 1ET100300517. Authors would like to thank to Ladislav Brom for painting Figure 1 and to Martin Majdloch to comments to English grammar.

References

1. Bakkum, J. D., Shkolnik, A. C., Ben-Ary G., Gamben P., DeMarse T. B., Potter S. M.: Removing some ‘A’ from AI: Embodied Cultured Networks. In: *Embodied Artificial Intelligence*, Springer 3139 (2004) 130-145
2. Beer, D. R.: The Dynamics of Active Categorical Perception in an Evolved Model Agent. In: *Adaptive Behaviour*, Vol. 11(4), SAGE Publications, London, UK (2003) 209-242
3. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. In: *Scientific American* (2001)
4. Brooks, A. R.: Intelligence without reason. In: *Proceedings of the 1991 International Joint Conference on Artificial Intelligence*, Sydney (1991) 569-595
5. Buriánek, F.: Karel Čapek. Československý spisovatel, Praha (1988)
6. Čapek, K.: R. U. R.. (Rossum’s Universal Robots). In: Čapek, K.: *Loupežník, R. U. R.., Bílá Nemoc*. Československý spisovatel, Praha (1983) 113-216 (in Czech)
7. Champandard, A.J.: *AI Game Development: Synthetic Creatures with learning and Reactive Behaviors*. New Riders, USA (2003)
8. The Global WordNet Association. Homepage: `http://www.globalwordnet.org/`
9. Havel, I. M.: Přirozené a umělé myšlení jako filozofický problem [Natural and Artificial Thinking as a philosophical problem]. In: Mařík, V., Štěpánková, O., Lažanský, J. et al.: *Umělá inteligence [AI] (3)*, **1**. Academia, Praha (2001) 17-75 (in Czech)
10. Lenat, D. B., Guha, R. V.: *Building large knowledge-based systems: representation and inference in the Cyc project*. Addision-Wesley (1990)
11. Lund H. H.: Co-evolving Control and Morphology with LEGO Robots. In: Hara, Pfeifer (eds.): *Morpho-functional Machines*, Springer-Verlag, Germany (2001)
12. McCulloch, W. S., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. In: *Bulletin of Mathematical Biophysics*, 5 (1943) 115-137
13. Minsky, M.: *The Society of Mind*. Simon and Schuster Inc. (1985)
14. Newell, A.: *Unified Theories of Cognition*. Harward University Press, USA (1990)
15. Ray, T. S.: An approach to the synthesis of life. In: Langton, C. G., Taylor, C., Rasmussen, S. (eds.): *Proceedings of Artificial Life, II*, Addision-Wesley (1991) 371-408
16. Russell, S. J., Norvig, P.: *Artificial Intelligence, A Modern Approach*. Prentice Hall, Englewood Cliffs, New Jersey (1995)
17. Sims, K.: Evolving 3d morphology and behaviour by competition. In: Brooks, R. A., Maes, P. (eds.): *Artificial Life IV: Proceedings of the Fourth International Workshop on the Synthesis and Simulation of Living Systems*. MIT Press, Ca, Massachusetts (1994)
18. Shaw, R.: The Agent-Environment Interface: Simon’s Indirect or Gibson’s Direct Coupling? In: *Ecological Psychology*, 15 (1), Lawrence Erlbaum Assoc. (2003) 37-106
19. Tanguy, E., Willis, P., Bryson, J.: A Layered Dynamic Emotion Representation for the Creation of Complex Facial Expressions. In: *Proceedings of the 4th International Workshop on Intelligent Agents, IVA 2003*. Kloster Irsee, Germany (2003)
20. Varela, F. J., Thompson E., Rosch, E.: The Embodied Mind. The MIT Press, Cambridge, Massachusetts, London, England (1991)
21. Weyns, D., van Dyke Parunak, H., Michel, F., Holvoet, T., Ferber, J.: Environment for Multiagent Systems: State-of-the-Art and Research Challenges. In: *Post-proceedings of the First International Workshop on Environments for Multiagent Systems*. LNAI 3374, Springer-Verlag, Germany (to appear)
22. W3C: Semantic web. Homepage: `http://www.w3.org/2001/sw/`
23. Wooldridge, M.: An Introduction to MultiAgent Systems. John Wiley & Sons (2002)
24. Yuret, D.: The binding roots of symbolic AI: a brief review of the Cyc project. Unpublished paper (1996)
25. Ziemke, T.: The ‘Environmental Puppeteer’ Revisited: A Connectionist Perspective on ‘Autonomy’. In: *Proceedings of the 6th European Workshop on Learning Robots (EWLR-6),* Brighton, UK (1997)