Towards Conscious Humanoid Robots

Jiří Wiedermann
Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 2, 182 07 Prague 8, Czech Republic
Email: jiri.wiedermann@cs.cas.cz

***Abstract*— We propose a simple yet cognitively powerful architecture of an embodied conscious agent. Our model differs from other proposals by exploiting two complementary internal world models. The first model captures the sensorimotor "syntax" of the agent's behavior and is used for situating the agent in its environment. The second model describes the sensorimotor dynamics of the world and is used for controlling the agent's behavior. Both internal world representations are fully determined by the agent's embodiment and its past experience. We show that the proposed model goes substantially beyond the potential of earlier models since it supports algorithmic processes underlying phenomena similar to higher cognitive functions such as imitation learning and the development of communication, language, thinking and consciousness.**

I. Introduction

The idea that non-trivial cognitive systems should build and exploit some form of internal world models has been around practically since the dawn of AI. However, efforts for controlling behavior by formal reasoning over symbolic internal world models have failed. Consequently, during nineteen nineties the mainstream research turned towards biology inspired behavior-based designs of cognitive systems. The respective approach has stressed the necessity of embodiment and situatedness used in sensorily driven control of behavior of simple robots (cf. [3]). This paradigm worked well with so–called subsumption architecture using incrementally upgraded layers of behavior realized by a task specific robot programming (cf. [11]). Nevertheless, after a series of promising successes, mostly in building various reaction–driven robots, it has appeared that such a framework has its limits. Especially in humanoid robotics a further progress towards higher levels of intelligence turned out to be impossible without introducing further innovations into the basic architecture of cognitive systems. This might be a reason for a present decay of engineering activities in the field of humanoid robotics compensated by a flush of theoretical works in the related fields, with machine consciousness being the ultimate goal (cf. [6]) .

Nowadays, it is generally believed that in order to break the before mentioned barrier reached by reaction-driven robots and to open the road towards higher brain functions we need automatic mechanisms that will augment the previously acquired knowledge. These mechanisms often make use of internal world models. Presently, prevailing trends seem to prefer other than symbolic representation of the internal worlds, in most cases variants of neural nets. For an overview of the recent state-of-the-art and a discussion on internal world models, cf. [7] or [4].

In [7], Holland and Goodman argue in favor of an internal world model consisting of two separated, but interacting parts: the agent-model and the environment-model. Recently, within the theoretical computer science a similar model has also been used by Blum et al. [2] in their quest for a formal definition of consciousness. In their work all the previously mentioned authors claim that a key to consciousness may be rooted in the formation and (co)operation of the two parts of their world models. Cruse arrives at a similar conclusion when considering an internal world model capturing the system's own body [4].

Our work builds on the works of previously mentioned authors. Our departure point will be the platform of computer science. As it is usual in software engineering, we will present informal functional specifications, in terms of the respective data processing requirements, of basic modules of a computational cognitive architecture. Then we give plausible arguments why we believe that the resulting computational model will support the realization of processes which mimic higher cognitive tasks such as imitation learning and development of communication, language, thinking and consciousness. Our cognitive architecture will make use of two cooperating internal world models. The first model is a so-called mirror net which learns frequently occurring "perception-behavioral units". These units are represented by multimodal information which is a fusion of sensory and motor information pertinent to a single "unit" of a situation. This part of the model is more or less static — once acquired a perception-behavioral unit remains stable. The design of a mirror net, which is responsible for the agent's situatedness in its environment, has been inspired by the assumed properties of the recently discovered mirror neurons (cf. [13], [12]).

In some sense, the mirror net represents both the environment and the agent; it captures both the syntax and semantics of a correct behavior. In the corresponding multimodal information the world is represented by sensory inputs while the corresponding agent's action by motor instructions and agent's "feelings" by proprioceptive feedback from the agent's internal sensors. Thus, on one hand, the mirror net captures similar information like the assumed internal world model (in fact, a neural net) featuring agent's body in the Cruse's theory or the agent-model in the Holland and Goodman's paper. On the other hand, since there are also environmental elements represented by the sensory information in the mirror net, in a certain fragmented way the mirror net also represents the environment, like Holland's and Goodman's environment-model.

The second part of our model is the agent's control unit. This unit receives multimodal information which is continuously delivered by the mirror net. The task of the control unit is to mine the knowledge from the flow of multimodal information. In the control unit the knowledge is represented by a (recurrent) network of concepts. The basic concepts correspond to units of multimodal information. However, the control unit also automatically computes the derived concepts which do not correspond to any existing multimodal information. These derived concepts correspond to the knowledge abstracted from the basic concepts. The control unit discovers, by statistical rules, the frequently occurring patterns in the flow of basic concepts, forms more abstract concepts and learns their time and space contiguity. The underlying network of concepts enables learning various patterns of behavior. Based on the current situation the control unit then determines the next action of the agent. The control unit can be implemented by an artificial recurrent neural network. Obviously, the control unit captures the dynamic aspects of the agent's interaction with its environment and has no any counterpart in either of the previously mentioned models by other authors.

Our model enables a plausible explanation of computational mechanisms underlying phenomena similar to higher brain functions inclusively that of consciousness. In our model, machine consciousness is a final phase of a sequence of increasingly more involved system's abilities in an increasingly stimulating environment. The respective chain of abilities starts with the ability to learn by imitation, goes through body, gesture and articulated communication among conspecifics and proceeds further via speaking to oneself until thinking ability is reached. Eventually, the previous development leads to a state that an entity possesses an ability to comment (or think of), in an abstract higher level language, any internal or external event, any past, present or expected phenomenon (*"to try on any 'story' for size"*, as Blum et al. have put it aptly [2]). In our model, this state is considered to be a hallmark of consciousness. This also corresponds well to Minsky's remark that "consciousness is a big suitcase" carrying many mental abilities [10].

The idea that the mirror neurons are at the heart of imitation learning ability and that they may play an important role in the development of the language, has been around since the beginning of this century (cf. [1], [8], [12]). One of the first computational models based on artificial mirror neurons has been described by the present author in [16]. In the present paper the idea of internal world models as prerequisites of machine consciousness and the respective mechanisms are further elaborated.

Summarizing, our results confirm in a constructive manner the intuition of the former researchers that suitable internal world models form the basis for the evolution of higher mental functions, inclusively those invoking consciousness.

The structure of the paper is as follows. In Section 2 we present our model in more detail. In Section 3 we describe its functionality leading to the emergence of computational consciousness.

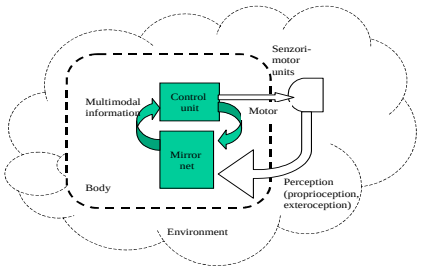

Fig. 1. The structure of a cognitive agent

II. The Model

The internal structure of our model is depicted in Fig. 1. It consists of four main parts: there are sensorimotor units, the sensorimotor world model represented by a mirror net, the control unit, and the body. Arrows depict the data flow between these parts. Next, we specify the actions performed by the model's individual parts. All data transferred along the arrows are of digital nature.

The *sensorimotor units* receive so-called *motor instructions* from the control unit. These are not only instructions for locomotive organs of the agent, but also instructions for pointing the sensors in a certain direction, for changing their settings, etc. At the same time, these instructions flow into the mirror net. The sensorimotor units deliver two kinds of data back to the mirror net.

The first kind of data is *exteroceptory data* that deliver information from the sensory units scanning the agent's environment. In this case, the sensory units act as a transformer of registered physical inputs (electromagnetic waves, sounds, pressure, etc.) into the digital form. In general, this transformation cannot be described mathematically since it depends on the physical/technical characteristics of the sensory units. The second kind of data is *proprioceptory data* delivering information from the internal sensors placed within the sensorimotor units or within the agent's body. For instance, this can be information about the current settings of the units or current conditions of the unit.

The next part of the model is the *mirror net.* It is a network of artificial mirror neurons which act analogously to (our ideas on) real mirror neurons. In each unit of this net (which might consists of several neurons), the exteroceptory and proprioceptory data from sensorimotor units meet with the motor instructions from the control unit and their conjunction is computed. This joint information is called *multimodal information.* The task of the mirror net is threefold:

Learning: the net learns frequently occurring multimodal information and stores its representation;

Identification: the net finds multimodal information already stored in the net which is "most similar" to the incoming information;

Associative retrieval: given only partial multimodal information in which the inputs from some sensorimotor units are missing, the net finds the entire multimodal information of which the partial information

is available.

In order to work in this way, we must establish that there is only a finite amount of "important" multimodal information stored in the mirror net; this can be achieved by a proper combination of "granularity" of perceptory data and finite increments in motor instructions. One can also consider some preprocessing of the information entering the mirror net, e.g., only "well separable" multimodal information is stored in the net, and to the incoming information its "nearest neighbor", in some sense, is sought. Analog or fuzzy neural nets seem to be attractive options for such purposes. The next requirement concerns the parts of the multimodal information. In order that the associative recall can work well, the entire multimodal information must be uniquely determined by any of its significant parts. For reasons that will be explained in the next section — namely in order the thinking mechanism to work — we assume that if there is a motor part in multimodal information, then this part alone determines the rest of multimodal information.

Each part of the mirror net specializes in learning and recognizing specific multimodal information corresponding to one "sensory-behavioral unit". Learning is done perpetually, when complete multimodal information appears at the input to the mirror net. Such circumstance is called *standard learning mode.* Learning proceeds by Hebbian principles, i.e., by strengthening the weights of neurons representing the respective multimodal information each time when it is recognized.

Thus, in any case, irrespectively whether all parts or only a (significant) part of the multimodal information enters the net, the net outputs complete multimodal information which proceeds into the control unit. In the context of the control unit, the representations of multimodal information are called the *concepts.* The task of the control unit is, given the current multimodal information represented by the active concepts, to produce a new set of active concepts. The motor part of multimodal information corresponding to these concepts is sent both to the sensorimotor units and to the mirror net. Clearly, the control unit determines the next action of an agent.

Within the control unit there are concepts corresponding to each occurrence of multimodal information in the mirror net. Moreover, new (abstract) concepts are formed within a control unit. Associations of various strengths connect the concepts within it. The concepts and the associations among them are all stored in the control unit and form the agent's memory. The rules of forming new concepts and strengthening the associations among them are based on the following principles; the first three of them have been identified already by Hume [9]:

> *contiguity in space:* two concepts get associated (or the respective association gets strengthened) if they frequently occur simultaneously; also, a new concept corresponding to the union of the two concepts gets formed;
>
> *contiguity in time:* two concepts get associated (or the respective association gets strengthened) if they frequently occur one after the other;
>
> *similarity:* a concept gets associated with another concept if the former is similar to the latter and vice versa; the notion of similarity must be appropriately defined (e.g., by requiring a sufficient overlap in multimodal information);
>
> *abstraction:* the common part of two similar concepts forms an abstraction of the two; the respective "abstract" concept is added to the concepts represented in the control unit.

The control unit should work according to the following rules. At each time, some concepts in it should be in active state. These concepts represent the current *"mental state"* of the agent. When new multimodal information enters the control unit it activates a new set of concepts. Based on the current mental state and the set of newly activated concepts a new set of concepts is activated. This set represents the new mental state of the agent and determines the next motor action of the unit.

Note that the new mental state is computed from an old one and from the new input. This mechanism reminds much the control mechanism in the finite automata. The idea is that the new mental state should be computable via associations stored among the concepts. In detail, the currently and newly activated concepts jointly excite, via the associations, a set of passive concepts. This excitation strengthens all the respective associations by a little amount. At the same time, small amount weakens the remaining associations. This models the process of forgetting. From among the set of all excited concepts, the set of the most excited concepts gets activated and the previously active concepts are deactivated. The set of currently active concepts is also strengthened. This set then represents the current mental state. The set of currently active concepts can be seen as the *short-term (operational) memory* of the agent. The set of all concepts with all settings of associations and weights can be seen as a *long-term memory* of the agent. Obviously, the control unit can also be implemented by an artificial neural net.

Based on the before mentioned principles the control unit is capable of solving simple cognitive tasks: learning *simultaneous occurrence* of concepts (by contiguity in space), their sequence, so-called *simple conditioning* (by contiguity in time), *similarity based behavior* and to compute their *abstractions.* In fact, these are the unit's basic operations. The mechanism is also capable to realize *Pavlovian conditioning* (cf. [14] p. 217), in which the control unit can be conditioned to produce a response to an apparently unrelated stimulus.

If one wants to go farther in the realization of the cognitive tasks, one should consider special concepts called *affects.* The affects come in two forms: positive and negative ones. The basic affects are activated directly from the sensors. Those corresponding to the positive feelings are positive whereas those corresponding to the negative feelings are negative. The associations can also arise among affects and concepts. The role of the affects is to modulate the excitation mechanism. With the help of affects, one can simulate the reinforcement

learning (so-called operant conditioning) and the delayed reinforcement learning. *Reinforcement learning* is learning where behavior is shaped and maintained by stimuli occurring after the responses rather than before. *Delayed reinforcement learning* is learning where the reinforcement stimulus — a reward or a punishment — does not necessary appear immediately after the step that will be reinforced. Pavlovian conditioning, reinforcement learning and delayed reinforcement learning seems to be a minimal test, which a cognitive system aspiring to produce a non-trivial behavior should pass.

In a stimulating environment during an agent's interaction with its environment concepts within the control unit start to self–organize, via property of similarity, into *clusters* whose centers are formed by abstract concepts. Moreover, by properties of time contiguity, chains of concepts, called *habits*, linked by associations start to form. The habits correspond to often performed activities. The behavior of agents governed by habits starts to prevail. In most cases such a behavior unfolds effortlessly. Only at the "crossings" of some habits an additional multimodal information from the mirror net (in an on–line or off–line mode — see the next section) is required directing the subsequent behavior. For more details concerning the work and cognitive abilities of the control unit, see the author's earlier paper [15] (and the references mentioned therein) where the control unit under the name "cogitoid" has been described.

The last component of our model is its body. Its purpose is to support the agent's sensorimotor units and to enclose all its parts into one protective envelope.

Now let us return to the question of internal models. Obviously, the mirror net can be seen as a specific kind of a static world model. In this model the world is represented in the way as it is cognised by an agent's sensory and motor actions, i.e., by an agent's interaction with its environment. It can be termed as *sensorimotor* model describing the "syntax" of the world. In the mirror net, the combinations of exteroceptory and proprioceptory inputs jointly with motor actions fitting together, which "make sense", are stored. Note that since proprioceptory information is always a part of multimodal information, also elements of an agent's own model are in fact available in the mirror net.

On the other hand, the control unit is a specific model of the world capturing the "semantics" of the world. In this model the relations among concepts are stored which, obviously, correspond to real relations among real objects and phenomena observed or generated by the agent during its existence. Similar relations are maintained also among the representations of these objects and phenomena. All this information represents a kind of a dynamic internal world model. One can also see this model as a depository of the "patterns of behavior which make sense in a given situation."

In the next section we describe how the interaction of both models leads to a more complex behavior.

III. Towards higher level cognitive functions

First we describe the mechanism of imitation learning which is a starting point for higher mental abilities (cf. [1], [8]). Imagine the following situation: agent A observes agent B performing a certain well distinguishable task. If A has in its repository of behavioral units multimodal information, which matches well the situation mediated by its sensors, then A's mirror net will identify the entire corresponding multimodal information (by virtue of associativity). At the same time, it will complement it by the flag saying *"this is not my own experience"* and deliver it to the central unit where it will be processed adequately. Thus, A has information to its disposal what B is about to do, and hence, it can forecast the future actions of B. "Forecasting" is done by following the associations in the control unit starting in the current mental state. Agent A can even reconstruct the "feelings" of B (via affects) since they are parts of the retrieved multimodal information. This might be called *empathy* in our model. Moreover, if we endow our agent by the ability to memorize short recent sequences of its mental states, than A can repeat the observed actions of B. This, of course, is called *imitation.*

The same mechanism helps to form a more detailed *model of self.* Namely, observing the activities of a similar agent from a distance helps the observer to "fill in" the gaps in its own dynamic internal world model (i.e., in the control unit), since from the beginning an observer only knows "what it feels like" if it perceives its own part of the body while doing the actions at hand. At this stage, we are close to *primitive communication* done with the help of *gestures.* Indicating some action via a characteristic gesture, an agent "broadcasts" visual information that is completed by the observer's associative memory mechanism to the complete multimodal information. That is, with the help of a single gesture complex information can be mediated. A gesture acts like an element of a higher-level (proto)language. By the way, here *computational emotions* can enter the game as a component of the communication. Their purpose is to modulate the agent's behavior. Of course, for such a purpose the agents must be appropriately equipped (e.g., by specific mimics, possibility of color changes, etc.). Once we have articulating agents, it is possible to complement and subsequently even substitute gestures by *articulated sounds.* It is the birth of a language. It is good to observe that the agents "understand" their gestures (language) via empathy in terms of their grounding in the same sensorimotorics, and in the more involved case, in the same patterns of behavior (habits), respectively (cf. [5]). One important remark: the transition from gestures to articulation does not only mean that gestures get associated with respective sounds but, above all, with the movements of speaking organs. Further, this facilitates still "speaking to oneself" and later the transition towards thinking (see in the sequel).

Having communication ability, an agent is close to thinking. In our model, *thinking means communication with oneself.* By communicating with oneself, an agent triggers the mechanism of discriminating between external stimuli (I listen what I

am talking) and the internal ones. This mechanism may be termed as *self-awareness* in our model. By a small modification (from the viewpoint of the agent's designer), one can achieve that the still self-communication can be arranged without the involvement of speaking organs at all. In this case, the respective instructions will not reach these organs; the instructions will merely proceed to the mirror net (see Fig. 1). Here they will invoke the same multimodal information as in the case when an agent directly hears the spoken language or perceives its gestures via proprioception (here we make use of our assumption that a motor part of multimodal information is sufficient to determine its rest). Obviously, while thinking an agent "switches off" any interaction with the external world (i.e., both perception and motor actions). Thus, in Fig. 1 do the dark parts of the schema depict an agent in a *"thinking mode";* this is captured by the cycle from the control unit to the mirror net and back to the control unit. In such a case, from the viewpoint of its internal mechanisms an agent operates as in the case of standard learning mode, i.e., when it receives the "real" perceptory information and executes all motor instructions. In the thinking mode, the same processes go on, but this time they are based on the virtual, rather than real, information mediated by the mirror net. One can say that in the thinking mode an agent works "off-line", while in the standard mode it works "on-line". Note that once an agent has the power of "shutting itself off" from the external world in the thinking mode, then this agent in fact distinguishes between a thought and reality.

In our model, we will informally define consciousness much in the spirit of Minsky's idea that "consciousness is a big suitcase", carrying many different mental abilities [10]. A prologue to consciousness is communication and thinking. The following "definition" of consciousness assumes that the agents are able to communicate in a higher-level language. A *higher-level language* is an "abstract" language in which a relatively complex action (corresponding to a sequence of mental states) or an abstract concept is substituted by a word expression or a gesture. A language level is the higher the "richer" the language is, i.e., the greater and more abstract is the set of things about which one can communicate. Agents can be thought of as being conscious, as long as their language ability has reached such a level that they are able:

- to speak or think on their own past, present and future experience, feelings, intentions and observed objects and actions and to explain their own behavior and expected phenomena;
- to imitate the observed activities of other agents, to speak or think on their (i.e., of other agents) past, present and future experience, feelings, and actions and to explain their observed or described behavior and intentions;
- to learn, and
- to realize activities given their verbal description in a high-level language.

Note that such a state of matters cannot be achieved without agents having an internal world model to their disposal along with the knowledge of the world's functioning and that of their own functioning within this world; to that end the agents must be constructed so that they can learn. A prerequisite for consciousness to emerge is an interaction among agents in a higher-level language with the same or similar semantics. Obviously, consciousness is not a property, which an entity does possess, or not. Rather, it is a continuous quality, which ranges from rudimentary forms towards the higher ones. And how do we know that a conscious agent "understands" its own actions and its world? Well — we simply ask it: the ability to answer such questions is the very idea of our "definition" of consciousness.

The above described notion of consciousness can be seen as a test to be applied to an entity in order to determine whether it is conscious according to that definition. Note, however, that we have brought arguments that a cognitive agent, designed in accordance with the proposed architecture, in principle could be conscious. From the functional and structural viewpoint, such an agent fulfills all assumptions needed for consciousness to emerge. It is a matter of an agent's proper embodiment, of appropriate technical parameters (memory capacity, operational speed, properties of sensorimotor units, etc.) of its modules, and of its suitable "education", whether in the agent consciousness will develop or not. The situation here is somewhat analogical to that in computing: any properly designed computer (obeying von Neumann architecture, say) is, in principle, a universal computer, but in order to do useful things it must be properly engineered and properly programmed. The same holds for our model with respect to thinking and consciousness. We believe that by our proposal we have made the first steps towards determining cognitive potential of a system not by testing the respective device but by inspecting its architecture.

ACKNOWLEDGMENT

This research was carried out within the institutional research plan AV0Z10300504 and partially supported by grant No. 1ET100300419.

REFERENCES

[1] M. A. Arbib: The Mirror System Hypothesis: How did protolanguage evolve? In: Maggie Tallerman, editor, Language Origins: Perspectives on Evolution. Oxford University Press, 2005.

[2] M. Blum, R. Williams, B. Juba, M. Humphrey: Toward a High-level Definition of Consciousness, Invited Talk to the Annual IEEE Computational Complexity Conference, San Jose CA, (June 2005)

[3] R.A. Brooks.: Intelligence without reason. Proceedings of the 12th Intl. Conference on Artificial Intelligence (IJCAI-91), 1991, pp. 569-595

[4] H. Cruse: The evolution of cognition — a hypothesis. Cognitive Science Vol. 27 No. 1, 2003, pp. 135-155

[5] S. Harnad: The Symbol Grounding Problem. Physica D 42: 335-346, 1990.

[6] O. Holland (Editor): Journal of Consciousness Studies, Special Issue: Machine Consciousenss. Volume 10, No. 4-5, April-May 2003

[7] O. Holland, R. Goodman: Robots With Internal Models: A Route to Machine Consciousness? Journal of Consciousness Studies Volume 10, No. 4-5, April-May 2003

[8] J. R. Hurford: Language beyond our grasp: what mirror neurons can, and cannot, do for language evolution. In: O. Kimbrough, U. Griebel, K. Plunkett (eds.): The Evolution of Communication systems: A Comparative Approach. The Viennna Series in Theoretical Biology, MIT Press Cambridge, MA, 2002

[9] D. Hume: Enquiry Concerning Human Understanding, in Enquiries concerning Human Understanding and concerning the Principles of Morals, edited by L. A. Selby-Bigge, 3rd edition revised by P. H. Nidditch, Oxford: Clarendon Press, 1975.
[10] M. Minsky: Consciousness is a big suitcase. EDGE, http://www.edge.org/ 3rd_culture/minsky/minsky_p2.html, 1998
[11] R. Pfeifer, C. Scheier: Understanding Intelligence. The MIT Press, Cambridge, Massachusetts, London, England, 1999, 697 s.
[12] V. S Ramachandran: Mirror neurons and imitation as the driving force behind "the great leap forward" in human evolution. EDGE: The third culture, http://www.edge.org/3rd_culture/ramachandran/ramachandran_p1.html, 2000
[13] G. Rizzolatti, L. Fadiga, V. Gallese, I. Fogassi: Premotor cortex and the recognition of motor actions. Cognitive Brain Research, 3:131-141,1996
[14] L. G. Valiant: Circuits of the Mind. Oxford University Press, New York, Oxford, 1994, 237 p.
[15] J. Wiedermann.: Towards Algorithmic Explanation of Mind Evolution and Functioning (Invited Talk). In: L. Brim, J. Gruska and J. Zlatuška (Eds.), Mathematical Foundations of Computer Science, Proc. of the 23-rd International Symposium (MFCS'98), Lecture Notes in Computer Science Vol. 1450, Springer Verlag, Berlin, 1998, pp. 152-166.
[16] J. Wiedermann: Mirror Neurons, Embodied Cognitive Agents and Imitation Learning. In: Computing and Informatics. Vol. 22, No. 6 (2003), pp. 545-559