HUGO: A Cognitive Architecture with an Incorporated World Model[1]

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

Abstract: *We present a design of cognitive system architecture with an internal world model. The internal world model is realized with the help of artificial mirror neurons. We consider generalized artificial mirror neurons acting both as a mechanism for assembling and learning multimodal sensorimotor information and as associative memory for invoking multimodal information given only some of its components. We show that within an artificial cognitive system a network of generalized mirror neurons can simultaneously serve as an internal world model recognized by the agent and as that of the agent's position within this world. We also specify a self-organizing control mechanism, which is based on the basic operations over concepts that were essentially identified by the British 18th century philosopher David Hume. This control mechanism makes use of the internal world model constructed in agent's interaction with real world and straightforwardly supports imitation learning. Building heavily on the properties of the generalized mirror net and on automatic abstract concept creation, we offer an algorithmic explanation of computational language acquisition, thinking and consciousness in our model. Rather than describing an implementation of the respective mechanisms, the aim of the paper is to present a plausible hypothesis concerning the architecture and functionality of artificial systems exhibiting higher cognitive functions.*

1. Introduction

Cognitive systems are instances of complex systems. In what follows, we will be interested in so-called cognitive architectures, i.e., in blueprints of artificial cognitive systems that take the form of embodied computers, or robots. Since robots are artifacts, they are designed by people to perform certain cognitive tasks. In order to fulfill these tasks both the control and the embodiment of a robot must be tailored to these tasks. In this paper, we will assume that each robot consists of a physical body holding the appropriate sensory and motor units, and of a control unit – a computer whose task is to learn to control and coordinate the available sensorimotor units. The goal of a robot's design is to produce a behavior that is qualified by robot's designers as a reasonable behavior realizing the cognitive tasks. We will neither be interested in the precise form of robot's embodiment nor

[1] This work was done within the Institutional Research Plan AV0Z10300504 and was partially supported by grant No. 1ET100300419

in its actual construction. Instead, we will be interested in the design of the overall robot's architecture, i.e., in specification of the robot's main modules, inclusively sensory, motor and control units, in specification of each module's tasks, and in the flow of information among these modules. To emphasize that the form of the embodiment will not be the main issue in our considerations, we will often speak about cognitive systems, or cognitive agents instead of robots. Our goal will be to devise a schema of a cognitive system, which will be as simple as possible, yet capable of offering a plausible algorithmic explanation of mechanisms that underlie cognitive functions usually termed as "higher brain functions". In other words, we will want to establish plausible arguments that our architecture will suit its task, i.e., that it will support the algorithmic realization of higher cognitive functions such as imitation learning, empathy, intentions, language acquisition, thinking and even consciousness.

In recent years we have seen a number of proposals of cognitive systems architectures, cf. [1], [12], [20], [22], [23] to name a few of them. All these proposals have aimed if not towards implementation, then at least towards explanation of higher mental functions. In fact, all of them were meant as proposals of experimental architectures with the goal to verify their viability in solving advanced cognitive tasks. In order to go beyond subsequently incremented subsumption architecture by a task specific robot programming, we need automatic mechanisms that will augment the previously acquired knowledge (cf. [23]). Moreover, an increased attention should be paid to the idea of internal world models, in which an agent could "simulate" alternatives of its future behavior [9], [20]. Recently, there was an attempt to formalize the notion of consciousness also in the field of theoretical computer science [4].

Our proposal also builds on the idea of a cognitive agent having an automatically constructed internal model of the world. This model captures that part of the environment in which the agent has situated itself (cf. [15] for the notion of situatedness). This internal world model includes not only the real world, which the agent has investigated during its life through its sensorimotor activities, but also a certain model of self as mediated by agent's exteroception and proprioception. Our model makes intensive use of known properties of mirror neurons. Since their discovery in nineteen nineties (cf. [17]) a great attention has been paid to mirror neurons by people interested in the theory of mind. Mirror neurons have also found their use in the cognitive models, mainly as a mechanism for sensorimotor coupling [8], [13]. Conjectures and theories have appeared pointing to a possible central role of mirror neurons in understanding the mechanisms of humanoid mind, e.g. language acquisition (cf. [1], [8], [16], [16]). Our model follows these lines of development. In our model, we have generalized the discovered ability of mirror neurons, viz. their activation, both when a specific motor activity is observed and performed. We suggest the artificial mirror neurons which work with arbitrary multimodal information composed of exteroceptory, proprioceptory and motor information. Such generalization is also supported by conclusions from recent neurobiological experiments (cf. [13]). The entire multimodal information is invoked also in cases when only a part of it is available. The importance of working with multimodal information has been stressed, e.g. in [5]. A net of the generalized mirror neurons is used to represent the frequently occurred "patterns" of multimodal information. The information represented in this net is shaped automatically,

in the process of agent's interaction with the world. The net of generalized mirror neurons supplies the complete multimodal information to a further control unit called controller. Its task is to compute the "next move" of the agent. A possible realization of the controller has been introduced by the author during the end of the nineteen nineties (cf. [25], [26], [27], [28]) under the name "cogitoid". Interestingly, the basic operations of a cogitoid, which works with concepts, have been inspired by the work of the English 18th century philosopher David Hume [10]. However, later it appeared that the full potential of the cogitoid could be exploited only when complementing it by the generalized mirror net. For the first time, the resulting model has been described in [29] mainly in the context of imitation learning. Since then, the idea about the cooperation of a controller with the generalized mirror net in understanding the humanoid mind has further crystallized. In particular, it has become obvious that in the model at hand the generalized mirror neurons have indeed played the role of agent's internal world representation. The current paper presents the summary of the author's current view of the respective problematic. It offers a relatively simple framework in which the algorithmic nature of higher brain functions can be explained. It presents the first step towards a concise theory of computational humanoid cognition, i.e., of non-trivial cognition in artificial entities. This theory is clearly inspired by biological reality, but it does not faithfully follow it in all aspects. This development is usual in any branch of human technology originally inspired by nature. By gaining more insight into the character of the respective processes, the technology eventually deviates from its natural template in order to make a good use of the already available achievements.

To simplify the references to our model we decided to call it HUGO. Unlike in other cases of "named" cognitive architectures, HUGO is not an acronym. Rather, this name refers to its Old German or Frank origin where it has meant "mind, thought, spirit, understanding, intelligence".

The structure of the paper is as follows. In Section 2 we present HUGO's blueprint and we specify the tasks of its individual modules and the flow of information among them. We concentrate to the functional specification of two essential modules: the net of generalized mirror neurons and the controller. We sketch how the net of the generalized mirror neurons is automatically formed, how the concepts in the controller develop and organize themselves and how the mirror net supports the working of the controller. In Section 3 we explain the mechanism of imitation learning, of empathy, of intentions, and based on this the evolution of inter-agent communication. From here, we go straightforwardly to the mechanism of thinking in HUGO and we sketch the conditions under which one might expect the development of consciousness. All our explanations will be based on specifications of individual HUGO's modules; we will not be concerned with details how these properties could be realized although we will occasionally suggest possible ways to do so. Thus, in analogy with the complex computational system design methodology used in the software engineering HUGO can be seen as the first phase of the design of such a system – namely its informal functional specification.

2. HUGO's architecture and functional specification

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

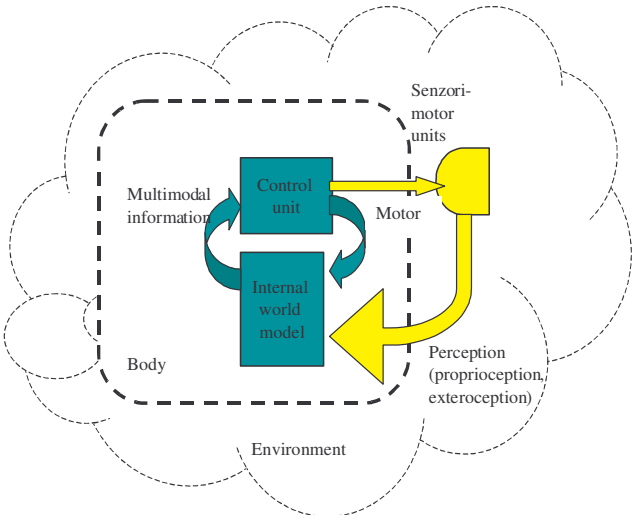

Fig.1: Embodied cognitive agent with an internal world model

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 agent's environment. In this case, the sensory units act as a transformer of analog inputs (images, electromagnetic waves, sounds, 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 is the mirror net. It is a network of artificial mirror neurons. In each unit of this net (which might consists of several neurons), the data from sensorimotor units meet with the instructions from the controller. 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 recall*: given only partial multimodal information in which the inputs from some sensorimotor unit 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. Fuzzy approaches seem to be attractive alternatives. 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, 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 "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 it is recognized each time.

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 called *controller*. In the context of controller, 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 the 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. Thus, 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*.

Between any pair of the concepts, there can simultaneously be both inhibitive and excitatory associations, or only one kind of them. 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 [10]:

- *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 cogitoid.

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 unite 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 the *long-term memory* of the agent.

Based on these 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, and to compute their abstractions. In fact, these are the unit's basic operations.

The control unit handles, e.g., *similarity-based behavior* in the following way. Assume that in the control unit there is an association between concepts A and B, denoted as $A \rightarrow B$, and that there is concept C similar to A: $A \sim C$. Now, if C becomes active, then C invokes A via similarity and A invokes B via simple conditioning. The mechanism is now capable to realize *Pavlovian conditioning* (cf. [22] p. 217), in which the control unit can be conditioned to produce a response R to an apparently unrelated stimulus A. The basic idea is first to learn via simple conditioning $S \rightarrow R$. Then, in the training process, we add a further stimulus A. Then, after some time, A alone invokes R (so-called cheating). In our model, this is because a simultaneous occurrence of S and A will cause the composed concept $A \vee S$ to form (by contiguity in space) and the conditioning will produce the behavior $A \vee S \rightarrow R$. Now, if only A will appear, $A \rightarrow A \vee S \rightarrow R$ will be realized (by similarity and conditioning). Interestingly, repeating the cheating many times A will cease to invoke S (so-called extinction). This is because in the absence of stimulus S accompanying A a new "parasitic" association $A \rightarrow X$ gets formed and strengthened (where X denotes the concept "nothing particular has happened"). Eventually, the association $A \rightarrow X$ prevails over $A \rightarrow A \vee S$ and thus the elicitation of R will cease. It can be shown that a

third phenomenon, called inhibition, is also within the reach of the controller designed along the previous lines (for details, see [27]).

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 arise also among affects and concepts.

The role of the affects is to modulate the excitation mechanism. An activation of a positive (negative) affect excites (inhibits) the associated concepts more than a standard (i.e. non-affecting) concept. Moreover, the concepts that get strongly associated with some affect "inherit" its quality - they also start to behave like affects as far as the excitation is concerned. 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 the minimal test, which a cognitive system aspiring to produce a non-trivial behavior should pass. As a matter of interest, it appears that after a suitable training a control unit designed in accordance with the previous principles, with sensorimotor units à la Turing machine tape and heads, can alone simulate any Turing machine [27]. The purpose of the training is to teach the system the transition function of the simulated Turing machine. In general, the controller is able to learn frequently repeated sequences of mental states, by connecting the respective concepts by chains of strong associations. Such sequences are also called *habits*. Habits can be seen as a rudimentary form of intentionality.

As far as the mechanism of abstraction is concerned, it is always on. Once a concept is activated, associations to similar concepts are automatically established and/or strengthened. New abstract concepts are formed via the relation of similarity, as the common part of the respective concepts. The "willingness" to create abstractions is controlled by the measure of similarity of the respective concepts and it could depend on the kind of overlapping multimodal information (motor, visual, etc.). In the memory of the controller, the concepts start to self-organize into so-called clusters. A *cluster* is a set of concepts that share a common abstraction, which lies in the center of the cluster. Note that via associations emerging due to similarity of concepts, the centers of clusters are activated always when any member of that cluster gets excited. This strengthens the presence of cluster centers. If a concept belongs to two or more clusters, all the respective centers are activated. The joint excitation of a concept from the activated clusters may activate that concept. This is how the activity in the controller changes from one mental state to the next one.

Obviously, all clusters are structurally similar - they are "made of" concepts and associations among them. However, they differ by their semantics. We can distinguish

three kinds of clusters: contextual, object, and functional ones. *Contextual clusters* evolve by a superposition (i.e., as abstractions via similarity relation) of concepts corresponding to multimodal information gained by perceptory units when observing the environment, i.e., the "context" in which an agent finds itself. Such concepts are also called *episodic memories*. In their centers, the abstract concepts characterizing the context prevail. Thus, a given context will excite the respective center, which abstracts, in sense, the most important features of the context. We see a form of *attentional mechanism*. Examples of contextual clusters are "on the street", "in the forest", "Christmas", etc. *Object clusters* evolve around specific objects. A respective object (or rather: its abstract representation) is in the center of the cluster and in the members of that cluster, the motor information prevails stating, what can be done with the object. Thus, the object clusters serve as a *role assignment mechanism* for objects. To select some concrete role, additional excitation from other concepts is needed. As an example, an object cluster named "key" can be mentioned, containing as its members contexts in which a key can be used, e.g. unlocking or locking a door, a safe, a car, etc. Finally, there are *functional clusters*. These are formed around frequently performed motor activities. A common abstraction of each of these activities presents the center of the respective cluster. In a sense, these clusters are "inverted" object clusters. They say, what for a motor activity is good. Thus, we can imagine functional clusters for unlocking a door, a safe, etc.

In a well-developed agent's memory, the chains of cluster centers forming habits govern the behavior of the agent at hand. Habits are present very strongly since they are continuously reinforced each time they are realized. As a result, an agent behaves as if seeking actively for opportunities to make use of habits that are appropriate to the given occasion.

We can conclude that the behavior of an agent is driven both by the chains of acquired associations as well as by the current context in which an agent finds itself. The current context activates similar, more abstract concepts that ``trigger" the respective behavior as dictated by the chain of the respective associations. Only occasionally, at "crossings" of some habits, an agent might enter a situation where an additional input, i.e. an additional excitation from other concepts is needed to direct the agent's further steps. The requirement to get additional input is also a part of the respective habit. Nevertheless, under similar circumstances, an agent with sufficiently evolved clusters and chains of associations will behave similarly as in the past. Even under a novel circumstance, chains of abstraction at higher levels will be found to "match" the current circumstance and drive agent's behavior. Thus in practice an agent can never find itself in a position when it does not "know" what to do. Note that in standard cases the agent's behavior will unfold effortlessly, without the necessity of making some additional "considerations".

As seen from the previous sketch, the algorithmic operation of a control unit tends to be quite involved. The experiments with simple controllers (cogitoids) have shown (cf. [3]) that the above-sketched mechanism seems to be very sensitive to settings of its various parameters (the excitatory/inhibitory increments, the weights of affects, etc.) and computationally quite demanding. Nevertheless, at present we are only interested in the specification of the control unit, rather than in its implementation. For the purpose of our

next explanations, the previous description should be sufficient. For earlier attempts to realize a controller in form of a cogitoid, cf. [25], [26], [27], [28].

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

3. Towards higher cognitive functions

In the previous section, we described mechanisms realizing the basic cognitive tasks in HUGO. Now we proceed towards cognitive tasks that are more complex. The first of them is imitation. Its working principle is simple, since the possibility of its straightforward realization has been the main reason for incorporating the internal model of the agent's world in form of the mirror net into the architecture.

At first sight, a mirror net does not look much as a model of the world. Nevertheless, in the mirror net "fragments" of the real world are indeed stored: its contents are in fact "episodic memories" consisting of multimodal complexes of related perceptions and actions as cognized by agent's perceptions and "verified" in practice by its motor activities. It is thanks to the learning mechanism of the mirror net that only those complexes are remembered here that have repeatedly shown up useful sometimes in the past. Note that since proprioceptory information with semantics, "*how it is like to perceive this and that stimuli and to perform this and that action*", is always present, also agent's own model is in fact available in the mirror net. Each piece of multimodal information represented in the mirror net describes "repository of atomic behavioral units" of the respective agent (cf. [13] for a similar observation). The agent's control unit assembles meaningful sequences from these pieces. These sequences govern the agent's behavior.

When it goes to imitation, 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 nets 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 knows what B is about to do, and hence, it can forecast the future actions of B. The "forecasting" is done by following the habit triggered by the current observation. Agent A can even reconstruct "feelings" of B, since they are parts of the recovered 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 *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 internal world model, since from the beginning an observer only knows "what it looks like" if it observes 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. By the way, here computational *emotions* can enter the game as a component of the communication. Their purpose is to modulate agent's behavior, similarly as was the purpose of affects. However, emotions are triggered by different mechanisms than affects. The latter are controlled "directly" by sensors, the former are controlled by activities of appropriate concepts (which, eventually, could be grounded in affects). 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 habits, respectively. One important remark: the transition from gestures to articulation does not only mean that gestures get associated with the 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). Note that our model explains, and thus supports, the classical linguistic hypothesis by Shapir-Whorf [19], [24], viz. language formation precedes thinking.

Having communication ability, an agent is close to thinking. In HUGO, thinking is nothing else than 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 controller to the mirror net and back to the controller. 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. According to Alfred Smee [21], an English physician of the 19th century, this is a hallmark of consciousness. George Dyson in [6] has suggested that this definition of consciousness has not yet been improved upon.

In our model, we will define consciousness yet as a higher-level mental faculty than Smee has envisaged, still subsuming his idea. Our approach is much in the spirit of Minsky's idea that "consciousness is a big suitcase", carrying many different mental abilities [14]. In our model, a prologue to consciousness is communication and thinking. The "definition" of consciousness assumes that the agents are able to communicate in a higher-level language. A higher-level language is not a language of motor commands (a machine language, speaking in the programming jargon). Rather, 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 world's functioning and that of their own functioning within this world; this state cannot be achieved without the agents being constructed so that they can learn. It is also good to realize that we do not require that the agents share the same construction principles (are of the same phenotype). What is the prerequisite for consciousness to emerge is an interaction among agents in a higher-level language with the same or similar semantics.

When speaking about speaking, thinking (which is almost the same in our model), and consciousness, a "standard" question arises: does an agent understand what it is speaking? In general, when speaking, an agent need not make any considerable effort. This is because we have indicated above that in our model, speaking is a (complex) habit, and as such, it must be learned. As explained in the previous section, habits follow association paths along more abstract concepts, which are centers of clusters. Some of these clusters represent "crossroads" where the speech production can branch. Various branches can be taken in accordance with the current mental state. In the standard mode there is no need to "understand" the language similarly as there is no need to "understand" one's acquired behavior. Much like the behavior, also speaking unfolds with ease, without the agent's permanent control, what it is speaking or thinking. The situation changes in the case of a conscious agent. Assume that we want a conscious agent to give a verbal explanation of a given word meaning. The word at hand, being an abstract notion, is a center of various clusters: e.g., of a context cluster, of an object cluster, of a functional cluster, and of other "unnamed" clusters. The members of each of these clusters in fact define this word. Giving the meaning of a word (or of an object named by that word) means to follow some paths

(associations) and to speak about them. Thus, contextual clusters describe situations in which the agent had encountered the given object described by that word; object clusters offer various roles which the respective object can play, and finally, functional clusters define the activities that can be related to (the objects denoted by) that word. Thus, in fact, "understanding" means to be able to generate many different short stories about the object of our interest, which are tailored to the agent's specific experience with that notion, or its prior knowledge about it. This ability follows from our definition of consciousness. The choice of a story depends on the circumstance the agent is in at this very moment (on its current sensory input, current emotions, association strengths settings, etc.). It seems that a similar conclusion also follows from [4]. In our scenario, the "proof of understanding" is given by the agent's explanation. Of course, instead of giving a verbal explanation an agent can merely think of the meaning of a word. In this way, he also becomes conscious of the meaning of that word.

In our model, consciousness is not a "yes" or "no" property, which an entity does possess, or not. Rather, it is a continuous quality, which ranges from rudimentary forms towards the higher ones, which are beyond our imagination since our human consciousness is not obviously its terminal instance. E.g., one can imagine consciousness endowed by a mechanism of exact recall of whatever we have seen, heard, felt, and experienced. On the other hand, it is also obvious that one cannot assume or "grow-up" a consciousness in too simple agents whose architecture or embodiment is too poor to handle, e.g., imitation.

It is interesting to observe that in spite of its relative simplicity our model in fact strengthens the Shapir-Whorf hypothesis: the language primary is not only in the development of thinking, it is even predominant in the development of consciousness. Only at a high level of abstract language development (and thus, of consciousness development) one can think of diminishing the dependence between mental development and embodiment and situatedness. Thus, we have reached the often-studied problem of a brain in the vat (cf. [17]). Such a brain could perhaps ponder on mathematical problems, but it will be devoid of any joy of life.

The above "definition" 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 the proper embodiment, of appropriate technical parameters (memory capacity, operational speed, properties of sensorimotor units, etc.) of its modules, and of suitable "education", whether 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 architecture 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 "opening it" and inspecting its architecture.

4. Conclusion

We have presented architecture (called HUGO) of an embodied cognitive agent with an incorporated world model. Its two main ingredients were the internal world model and the controller. The internal model of the world was based on the generalized idea of mirror neurons. The controller was based on the idea of self-organizing memory in which new concepts and association among them form automatically. We have presented "functional specifications" of these two units, which were sufficient to give plausible explanation of the evolution of higher cognitive faculties in our model, such as imitation learning, empathy, communication, thinking and eventually consciousness. HUGO seems to be one of the first cognitive models, which is able to explain the respective cognitive phenomena to such an extent. It represents a bridge between the theory of mind and the computational models of mind. So far, this model presents a hypothesis describing the algorithmic nature of higher cognitive tasks; it is but the first step towards realization of genuine cognitive systems exhibiting interesting cognitive behavior. At the same time, the model represents an attempt to characterize cognitive systems possessing advanced cognitive abilities "structurally", by their architecture. In order to advance along these lines, a further development of the model, based on experiments aimed at validation of ideas presented in this paper, is needed.

References

[1] M. Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y.: An integrated theory of the mind. Psychological Review 111, (4) pp. 1036-1060, 2004.

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

[3] M. Beran: Cogitoid: from Reflexes to a Smart Bacterium (in Czech: Kogitoid - od reflexu k chytré bakterii) In: Proceedings of the 2nd workshop Cognition and Artificial Life II, Silesian University, Opava, 2002, ISBN 80-7248-151-7

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

[5] B. Chandrasekaran: Multimodal Representations as Basis for Cognitive Architecture: Making Perception More Central to Intelligent Behavior, in Intelligent Information Processing, Mark Musen, Bernd Neumann, & Rudi Studer, Eds., IFIP International Federation for Information Processing Series, Vol. 93. Dordrecht: Kluwer Academic Publishers, pp: 13-16, 2002

[6] G. Dyson: Darwin Among The Machines: The Evolution of Global Intelligence. Reading: Addison-Wesley, 1997.

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

[8] M. Haruno, Wolpert, D. M., & Kawato, M. (2001). MOSAIC Model for sensorimotor learning and control. Neural Computation, 13, 2201-2220

[9] O. Holland: The Future of Embodied Artificial Intelligence: Machine Consciousness?. In: F. Iida et.al. (Eds.): Embodied Artificial Intelligence, International Seminar, Dagstuhl Castle, Germany, July 7-11, 2003, Revised Papers. LNCS, Springer 2004, pp. 37-53
[10] 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
[11] 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.
[12] P. Langley: An adaptive architecture for physical agents. Proceedings of the 2005 IEEE/WIC/ACM International Conference on Intelligent Agent Technology (pp. 18-25). Compiegne, France: IEEE Computer Society Press, 2005.
[13] G. Metta**,** G. Sandini**,** L. Natale**,** L. Craighero and L. Fadiga: "Understanding mirror neurons: A bio-robotic approach". Epigenetic robotics, Metta**,** Giorgio and Luc Berthouze (eds.), J. Benjamins Publ. Co., pp. 197—231, 2006
[14] M. Minsky: Consciousness is a big suitcase. EDGE, http://www.edge.org/3rd_culture/ minsky/ minsky_p2.html
[15] R. Pfeifer, C. Scheier: Understanding Intelligence. The MIT Press, Cambridge, Massachusetts, London, England, 1999, 697 p.
[16] 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
[17] V. S. Ramachandran: Mirror neurons and the brain in the vat. http://www.edge.org/ 3rd_culture/ ramachandran06/ramachandran06_index.html
[18] G. Rizzolatti, L. Fadiga, V. Gallese, I. Fogassi: Premotor cortex and the recognition of motor actions. Cognitive Brain Research, 3:131-141,1996
[19] E. Sapir: 'The Status of Linguistics as a Science', 1929. In E. Sapir: Culture, Language and Personality (ed. D. G. Mandelbaum). Berkeley, CA: University of California Press, 1959
[20] M. P. Shanahan: Consciousness, Emotion, and Imagination: A Brain-Inspired Architecture for Cognitive Robotics, Proceedings AISB 2005 Symposium on Next Generation Approaches to Machine Consciousness, pp. 26-35, 2005
[21] A. Smee: Principles of the Human Mind Deduced from Physical Laws, 1849 (N.Y. 1853)
[22] L.G. Valiant: Circuits of the Mind. Oxford University Press, New York, Oxford, 1994, 237 p.
[23] J. Weng, Y. Zhang: Developmental Robots - A New Paradigm. In Prince, Christopher G. and Demiris, Yiannis and Marom, Yuval and Kozima, Hideki and Balkenius, Christian, Eds. Proceedings Second International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems 94, pp. 163-174, Edinburgh, Scotland, 2002
[24] B. L. Whorf: 'Science and Linguistics', Technology Review **42**(6): 229-31, 247-8, 1940. Also in B. L. Whorf: Language, Thought and Reality (ed. J. B. Carroll). Cambridge, MA: MIT Press, 1956

[25] J. Wiedermann: The Cogitoid: A Computational Model of Mind. Technical Report No. V-685, Institute of Computer Science, Prague, September 1996, 17p.
[26] J. Wiedermann, J.: Towards Computational Models of the Brain: Getting Started. Neural Networks World, Vol 7., No.1, 1997, pp. 89-120
[27] J. Wiedermann, J.: The Cogitoid: A Computational Model of Cognitive Behaviour (Revised Version). Institute of Computer Science, Prague, Technical Report V-743, 1998, 17 p.
[28] 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.
[29] J. Wiedermann: Mirror Neurons, Embodied Cognitive Agents and Imitation Learning. In: Computing and Informatics. Vol. 22, No. 6 (2003), p. 545-559