Designing A Self-Aware Neuromorphic Hybrid

Alexei V. Samsonovich[1] and Kenneth A. De Jong[2]

[1,2]Krasnow Institute for Advanced Study
[2]Computer Science Department
[1,2]George Mason University, Fairfax, VA 22030-4444
[1]asamsono@gmu.edu, [2]kdejong@gmu.edu

Abstract

A top-level integration of symbolic and connectionist components of a cognitive architecture can be achieved through the blending of innovative concepts used in both components. i) On the connectionist side the key innovative elements are neuromorphic cognitive maps that provide for associative indexing and organization of symbolic memories and path-finding in modeled cognitive spaces. On the symbolic side the key innovative elements are: ii) a unique, central notion of "self", and iii) a formal representation system based on the general notion of a "schema". In this work we describe a theoretical framework allowing us to design an implementable cognitive architecture of this sort.

The Challenge of Integration

The grand vision of integrated, versatile, intelligent systems capable of unlimited, autonomous cognitive growth starting from a minimal embryo [22] is a difficult, but probably the most interesting and critical challenge of our time. Achieving this meta-goal of artificial intelligence requires certain key features of human cognition to be enabled in an agent: (a) the cognitive growth ability consisting of basic human kinds of learning: abilities to learn problem spaces autonomously (by exploration of environments and previously acquired episodic memories) and from an instructor (by consultation or by mimicking instructor's behavior); (b) basic human kinds of memory, including episodic memory understood as memory of personal experience of the subject; (c) social meta-cognition (simulationist 'theory-of-mind' capabilities: can simulate other minds), including social imagery ('what if' capabilities: can imagine self and others in plausible situations); (d) real-world communication capabilities, including anaphoric reference resolution and other techniques studied by discourse representation theory [41] that use mental simulations; (e) higher emotional capabilities, including appreciation of subjective feelings resulting from emotional evaluation of mental states and the ability to understand and to simulate emotions observed in other agents.

It is well known that in humans the development of (a) – (e) critically depends on (1) the neuronal networks of the brain and (2) "*the child's conception of self during the age-4 transition*" [4]: this transition, according to Barresi [4] and to other works in developmental psychology, consists in rapid, simultaneous development of the entire complex of those cognitive abilities (a) – (e) that require the Self as a precondition. Therefore, mimicking a human self in a neuromorphic artifact could be a major step toward enabling the emergence of (a) – (e) upon implementation.

A straightforward approach to mimicking the human self in a neuromorphic cognitive system could be to study the brain, to describe and to replicate its functionality step by step. The progress made in this direction in brain sciences during last decades is tremendous: brain areas related to highest cognition in humans are identified (such as the hippocampal formation, medial temporal and frontal lobes), their functions are characterized, connectionist models of the underlying mechanisms are studied numerically and related to brain imaging. At the same time, there is an enormous gap between brain sciences and artificial intelligence. We believe in a possibility to bridge this gap based on a mapping between the brain architecture and cognitive architectures, expecting the complex of structure-function relations be capable of self-completion in a new substratum. It is therefore wise to use results from cognitive and computational neuroscience in our self-aware cognitive architecture design that logically takes off from studies of the human mind and the self.

The Proposed Approach

The Self concept

According to our view, in order to be self-aware, an agent must have an image of its Self consistent with self axioms (see below), must attribute to this Self (as to the subject) all first-hand experiences represented in the cognitive system, and must believe that this Self is responsible for all behavioral and cognitive actions initiated by the agent.

The specific detail that separates our approach from others is our notion of the Self: it does not refer to the actual physical (hardware) or informational (software) body of the agent. The Self in our framework is an imaginary abstraction [28, 33] rather than the agent per se. Nevertheless, this imaginary abstraction has real representations in the cognitive system that are maintained over time and used to guide cognitive information processing. The imaginary Self is represented as if it was a real entity, thereby forcing the system to behave as if it had this real entity inside. This is the first of a set of fundamental principles called here self axioms that are permanently built into our cognitive system at birth, as we claim they are permanently built into the human mind [33].

Self axioms are beliefs of the agent about own self. They are implemented as constraints on the semantics of possible representations in the system and together enable the kind of behavior and cognitive abilities that are consistent with the common-sense notion of a conscious self. The entire set of self axioms is described in detail in [33], and summarized here: (i) The Self is a unit that exists as an abstract entity associated with a specific cognitive system. (ii) This unit is the only subject of all first-hand experiences represented in the system. (iii) This unit is the only author of all self-initiated actions of the system. (iv) This unit is unique, one and the same over time, in all circumstances. (v) This unit is indivisible. It has no internal parts or substructure. (vi) This unit is self-consistent over time. (vii) This unit is capable of acting independently. (viii) This unit is always localized in the agent's body and in time. (ix) This unit is capable of self-awareness. Again, (i) – (ix) are beliefs rather than facts.

Implementation of these principles in the brain is based on (i) association of represented experiences with instances of the Self (self-attribution), (ii) dynamic rules sensitive to this association; (iii) semantic constraints on schemas. In long term memory, (i) is subserved by the contextual cognitive map (the hippocampus: [25]) that provides unique identifiers for all logically meaningful instances of the subject's Self associated with remembered experiences. The result is called episodic memory, i.e., memory of personal experience [37-39]. In working memory, the function of self-attribution is subserved by the frontal lobe.

The self in our model has discrete instances (e.g., "I" taken at different moments of time) that are represented by simple tokens labeled *I-Now*, *I-Previous*, *I-Next*, etc. Accordingly, all representations of experience are partitioned into mental states attributed to those tokens (instances of the Self). A given mental state represents a particular mental perspective of the Self, together with all experiences attributed to this instance of the Self. A self-aware mental state is a mental state that is aware of the instance of the Self to whom it is attributed. This element of awareness is represented by a token "me-now" included in the content of the mental state (in addition to the label *I-Now*). A mental state may be aware of other mental states, indexing them by similar tokens: "me-next", "me-previous", etc. These tokens-references are used to control information exchange between mental states. As a result, active mental states form a dynamical lattice and evolve in parallel, interacting with each other. These interactions are constrained by self axioms. An essential feature enabled by the lattice of mental states is that this framework allows the system to process each mental state from another mental state (mental perspective), thereby providing a basis for various forms of meta-cognition [33]. Active mental states constitute working memory. Episodic memory includes frozen mental states that once were active.

Biologically, mental state tokens can be realized as stored activity patterns in the hippocampus (for episodic memory) and spatio-temporal patterns or phases of neuronal firing in the neocortex for working memory. The two kinds of representations are presumably linked to each other via the phenomenon known as the phase precession [16]: as a result, the number of mental states simultaneously active in the brain is limited to 7±2 [18].

The schema formalism

A set of advanced symbolic cognitive architectures based on production systems were developed in artificial intelligence intended as models of the human mind. The main of them are Soar [15, 22], ACT-R [1, 2], EPIC [17], and Polyscheme [8]. The base of our cognitive architecture is a framework of schemas that can be viewed as a generalized production system that offers, on the one hand, a meta-level of information processing, as compared to Soar, Act-R and Epic, and on the other hand, incorporation of non-symbolic primitives into schemas (cf. [8]). Our schema is a building block generalizing productions, operators and chunks. Here the term "schema" refers to an abstract model or a template that can be used at any level of abstraction to instantiate and to process a certain class of mental categories [33]. Schemas may include concepts, elementary actions and their sequences (scripts), as well as templates for instantiations of qualia and abstract notions.

Generally, a schema can be represented as a graph, in which external nodes are called terminals, the first of which is called the head and represents the instantiated category of the schema. A simple example is a schema of the number two that has three terminals: the head signifying that there are two entities in a given context and other two terminals that can bind to any entities. This schema has no internal nodes. More complex examples can also be given (Figure 2). The first step of a state creation based on a schema is called binding. Links of the graph specify how bindings of internal nodes to other internal nodes and to terminals should be established. Bindings of terminals to external content are established based on their attributes. The notion of an attribute used here is a meta-level as compared to that in Soar and will be described elsewhere.

A remote analogy between our schemas and LISP functions can be seen. Consistently with the analogy, our schemas can create other schemas, thereby enabling the cognitive growth of the system. This is the main distinctive feature of our system of schemas: qualitatively new cognitive abilities represented by new schemas can be created autonomously by existing schemas. In this sense, schemas are divided into innate (pre-programmed) and acquired (automatically created by the system). In this case, evolution of the set of schemas can be controlled using paradigms of genetic programming [14, 31, 3] with an appropriate definition of fitness. Another similarity with

LISP is in that internal nodes in schemas are references to other schemas or to *primitives* (non-symbolic or lower-level symbolic tools and routines designed to perform specific functions, e.g., sensory signal processing, motor command execution, mathematical calculations). Primitives constitute procedural memory, while the set of schemas constitutes semantic memory in our cognitive architecture. A multi-plane hierarchy of semantic memory follows from the above: (1) complex schemas contain simple schemas as their components, (2) schemas representing specific concepts are clustered into general categories represented by more abstract schemas, and (3) acquired schemas refer to their innate prototypes.

In the brain, higher schemas are known to be implemented in the medial temporal lobe and in the prefrontal cortex as stored spatio-temporal patterns of neuronal activity. Biological neuronal networks provide the capacity for associative schema mapping in creation of cognitive representations. We use the same neuromorphic principle in our approach: associative retrieval of an appropriate schema is one of the functions of our conceptual cognitive map.

Cognitive architecture mapped onto the brain

Here is a bird-eye view of our proposed architecture (Figure 1). The core of it includes the four memory systems: working, episodic, semantic (representations in which are based on schemas) and procedural (primitives). The input-output buffer operates in terms of states of schemas and interacts with working memory. The driving engine and the reward-punishment system "run" the above components. They are implemented algorithmically. Cognitive map is an artificial neural network playing the central role.

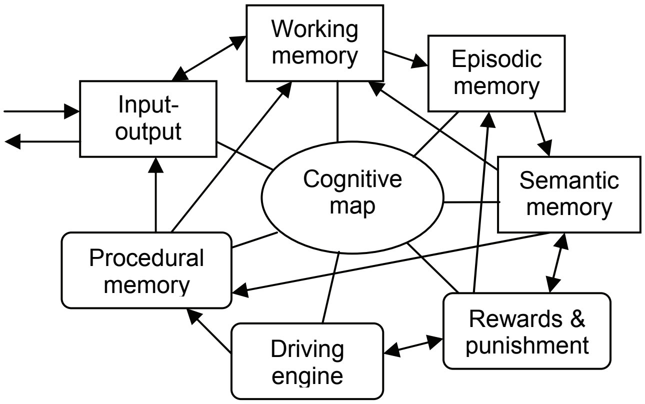

Figure 1. The hybrid cognitive architecture at a large scale. Shapes reflect the nature of components. Rectangle: higher-level symbolic, oval: analog connectionist, rounded rectangle: algorithmic. Arrows indicate essential data flow.

Our complete cognitive architecture at the level of its components, their parts and cognitive functions can be mapped onto structures and systems of the brain. Specifically, the mapping should cover structures, systems and features of the human brain listed below (and definitely more), where a separate list is provided for each component of the architecture shown in Figure 1 (e.g., "our working memory" is the architectural component, not the working memory system of the brain).

- ❖ Our working memory (active mental states): neuronal activity in virtually all non-primary neocortical areas (excluding somatosensory and olfactory modalities) and related thalamocortical loops.
- ❖ Our episodic memory (frozen mental states): synaptic efficacies in the same areas as listed above, plus the hippocampus (providing allocentric mental state labels) and the medial temporal lobe (mediating hippocampo-cortical connections).
- ❖ Our semantic memory (schemas): medial temporal lobe and other higher neocortical areas.
- ❖ Our input-output buffer (sensory and motor cognitive states): neuronal activity in premotor, motor, auditory and visual neocortices, and in related structures of their thalamocortical loops.
- ❖ Our procedural memory (primitives): stored patterns in specialized neocortices and related structures, including visual, auditory, speech (Broca, Wernike), motor and premotor areas and the cerebellum.
- ❖ Our reward-punishment system: dopaminergic system, including parts of the frontal lobe, basal ganglia (e.g. nucleus accumbens, ventral pallidum), most of the limbic system, substantia nigra.
- ❖ Our driving engine: in general, there is no unique brain system corresponding to this component. In our architecture it is an algorithmic shell running all other components (including all memory systems) as passive data structures. Specific functions of synchronization and timing can be associated with the cerebellum, the pons and nuclei of the thalamus.
- ❖ Our cognitive map has three components: contextual, conceptual and emotional. Related brain structures include the limbic lobe, the amygdala, the hypothalamus, the medial temporal lobe cortices, the cingulate, orbitofrontal and parts of the parietal and prefrontal cortices.

Cognitive map is more than the 'virtual hippocampus' in our architecture

Our cognitive map component is an associative neural network that is primarily inspired by the hippocampus and its role in the brain. In addition, it performs certain cognitive functions of other structures (see above). Therefore, the functions of our cognitive map extend beyond a 'virtual hippocampus'. In our architecture (Figure 1) this component is partitioned into contextual, conceptual and emotional maps and is used for integration of episodic, semantic and working memory systems with each other and with other components, subserving the following functions.

(1) Associative indexing of all stored memories. In particular, this function enables cued recall: retrieval of a

memory by associative pattern completion, based on partial information about an episode or a concept [35, 9].

(2) Modeling multidimensional cognitive spaces and path integration. Associative neural networks can learn continuous attractors with arbitrary topology and geometry that can serve as internal cognitive models of problem spaces [32]. In our architecture, this function is involved in memory management (comparison, hierarchical organization, clustering of memories), in generation of emotional feelings (see below), and in analogy search within a certain context or a certain domain of knowledge based on global characteristics of a remembered episode or a concept (proximity of pointers: see above).

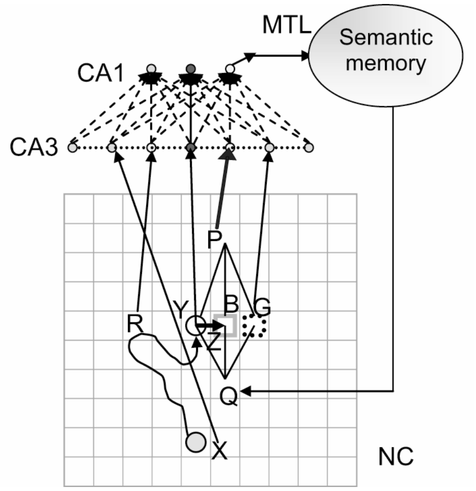

Figure 2. Integration of neuromorphic and symbolic components. Top: neural network architecture of the cognitive map component [27]. MTL: medial temporal lobe, NC: neocortex, CA1 and CA3: fields in the hippocampus proper. Bottom: incremental learning in a push-push environment. The agent body (filled circle) initially located at X has to push the block B (gray square) to the location G (dotted square). To do this, the agent gets to Y using an available schema R. Schema P applied to episodic memory of a successful trial recognizes that Y is opposite of G with respect to B. Schema Q recognizes that the pushing move Z is directed toward the target G. Together P and Q provide a basis for a new schema of a one-step push, that will be tested in novel situations. Interface of symbolic and connectionist components (indicated by solid arrows) is mediated by MTL.

(3) Strategic retrieval: path-finding in memory space ([27]: Figure 2). In our architecture, this mechanism is involved in context reinstatement during episodic retrieval, in assembling critical elements during new schema formation, and in intuitive planning and problem solving (e.g., spatial navigation) based on cognitive space search.

(4) Multiple trace formation (described by the multiple trace theory: [19, 20, 21]). This function is involved in memory re-consolidation during off-line replay, including reinterpretation and compilation of previously acquired memories (see below).

The above functions can be achieved based on the well-known properties of associative neural networks, such as their ability to learn point attractors and feature maps (for a reference see [12]), as well as their ability to store and to navigate graphs and manifolds of arbitrary dimensionality [27, 32]. Our confidence in success of reproduction of (1) – (4) using the cognitive map component is based on the results of (i) brain imaging studies, (ii) parallel multiunit recordings in freely behaving animals, and (iii) connectionist modeling in cognitive psychology, including our own previous results. For example, our neural network model of the hippocampus (Figure 2, top: see [27] for details) solves complex problems of spatial and non-spatial navigation, is robust up to realistic mammalian connectivity, and is ideally suited to navigate the space of episodic memories in the proposed cognitive architecture.

The three components of our cognitive map serve the three symbolic memory systems: episodic, semantic and working. Episodic cognitive map provides a model of an abstract space indexing remembered episodes. Semantic cognitive map provides a model of an abstract cognitive space indexing available schemas and schema categories. Emotional cognitive map provides a model of the space of possible emotions, feelings and affects and its associative mapping onto possible activity conditions and contents that may be active in working memory.

Cognitive Architecture in Action

Contextual reinstatement: An example of a routine task solved by the proposed architecture

The following sequence could be a routine procedure executed every time when a new session of the agent is started. As the system "wakes up", a new mental state *I-Now* is initiated representing the current instance of the self of the system. Its initially empty content is populated by states representing sensory input (copied from the input-output buffer). In parallel, contextual attributes of the token *I-Now* (that specify "who I am", "where I am", "what time it is", "am I awake or dreaming", etc.) are reinstated based on the relevant semantic knowledge and on the current experience. This process of reinstatement is controlled by the cognitive map, as described in [27]. Simultaneously, a mental state labeled *I-Goal* is retrieved from prospective episodic memories ("what am I supposed to do today"). The most recent retrospective episodic memory ("what I was thinking before falling asleep") is reactivated as I-Previous. The content of *I-Now* starts evolving, resulting in ideas (here "an idea" is understood as a state representing a feasible action) produced by mapping of action schemas. The three mental states (*I-Previous*, *I-Now*, *I-Goal*) get represented in a newly created meta-cognitive mental state *I-Meta*, where they form a working scenario. The next event is the selection of immediate intention based on available ideas in *I-Now* and their consistency with the working scenario. When this happens, the mental state *I-Next* is initiated, and its content is populated with the scheduled action and the expected outcome (e.g., "I am getting up"). At this point, the dynamical lattice of active mental states already looks similar to Figure 2. At the next moment of time the agent performs the intended action and simultaneously shifts the perspectives of its mental states. *I-Next* takes the position of *I-Now*, *I-Now* becomes *I-Previous*, *I-Previous* becomes *I-Yesterday*, is deactivated and returns to episodic memory.

Subsequent evolution of the mental state lattice may vary. For example, a new instance of *I-Next* may be initiated based on new ideas and intentions. Alternatively, *I-Meta* may take the place of *I-Now* (with *I-Goal* shifting to *I-Next*, etc.), and the agent would start reasoning about own plans from a greater perspective. In another possible scenario, a new mental state *I-Imagine* is created to explore some interesting possibility by mental simulation. While specific paradigms of (meta)cognition are innumerous, normally the state of working memory may include a meta-level perception of own mental states (*I-Meta*), a notion of the current goal (*I-Goal*), a sense of what is going to happen next (*I-Next*), a sense of what just happened (*I-Previous*), and the sense of the current, actual instance of the subject's mind (*I-Now*).

The Self provides leverage for cognitive growth

Schemas can be created automatically by various mechanisms. E.g., the following strategy can be implemented as an innate schema and used by the agent for incremental learning in a broad variety of exploratory paradigms that allow for multiple trials. Suppose that, given a certain initial situation *X* in a virtual world *W*, the task for the agent is to achieve a certain result *G*. The meta-task is then to learn how to do this in the world *W* in general. The initial strategy of the agent (the first step) is to perform randomly generated sequences of moves, every time starting from *X*. Suppose that in a number of trials, *G* was achieved several times. The next step is to discover the mechanism of this event. To do this, the agent starts by building a theory of what happened. Available matching schemas are applied to the remembered sequence of actions and other events that led the agent from *X* to *G*, until all intermediate steps and critical events are "understood" (i.e. mapped by schemas).

The key in this process is self-attribution and reinterpretation of episodic memories: an element that requires the self concept. Specifically, the system finds "motivations" of own actions (that in fact were randomly generated) and puts them together in a new schema, creating a belief that this schema was actually used in the remembered episode. In addition, the system automatically ensures that self-axioms apply to this new memory interpretation; e.g., the remembered experiences and actions are consistent with each other. At the next step, the agent tests the hypothesis that the new schema works in general. This is done by applying the schema to new encountered situations. Based on the tests, the agent may reject, revise or accept the schema. When accepted, the new schema becomes a building block in further exploration and learning. This example therefore demonstrates the leverage of a self-concept in incremental learning [29, 30].

Having a human-like Self becomes critical in paradigms like simulated intrusion detection

Here we consider a paradigm in which an agent learns from an instructor that there is a world called "simulated intrusion detection game", where the agent is a network administrator of a computer grid shown in Figure 3, and its job is to protect this grid from a hacker. According to the rules, the agent may set any 5 out of 24 internal computer nodes to a high security level (which would require a double time for a hacker to gain root access, as compared to "easy" nodes) and make any one of the rest a honeypot (where any intrusion will be instantly detected). Detecting an intrusion will take a finite time elsewhere. An intrusion may happen spontaneously at any time. A hacker enters through the gateway, can "see" neighboring nodes, including their levels of security, and can crack and occupy any of them. Its goal is to get access to a maximal number of nodes before being detected. An additional rule is that the agent itself is allowed to attack its own grid, pretending to be a hacker, in order to learn how efficient its defense is.

Suppose the agent starts by setting the grid at random and then attacking it, pretending to be a hacker. The agent will use its self concept to simulate a hacker: specifically, it will use the notions of a goal, a voluntary action, a working scenario, etc. Its first model of a hacker (inferred from episodic memories left after the pretend-play) will be a naïve hacker that simply prefers easy nodes to hard nodes. Based on this model and its mental simulations, the agent will now design its first defense scheme, with hard nodes arranged into a "herding" pattern that will lead a hacker to the honeypot (Figure 3). In the next set of experiments, a simulated hacker will take into account a possibility of a honeypot (that looks like an easy node), the Self of the network administrator (the agent) and its intention to capture the hacker into the honeypot. As a result, this hacker will be unlikely to follow the path shown in Figure 3. Therefore, the second model of a hacker created by the agent will be a "honeypot-aware hacker" that will exhibit a completely different behavioral pattern. The process will continue, converging to an optimal solution.

It is true that the same or better result could be achieved based on game theory, traditional machine learning or evolutionary computation. It is also a fact that these traditional approaches require a certain degree of translation of the problem space into a computer code done by a human a priori. The approach proposed here is free of this requirement; instead, it requires the notion of a self in the agent, and has certain advantages when something unexpected happens. For example, during one of the agent's experiments, the grid may be attacked by a real hacker. In order to be able to discriminate between a real hacker and a simulated hacker in this situation, the agent must process at least two mental states in parallel: one, *I-Pretend*, which is simulating a hacker, and another, *I-Real*, that will monitor all system, will be aware of the current activity of *I-Pretend*, and therefore will be able to detect the hacker. The bottom line is that at least two mental states must be processed in parallel in examples like this.

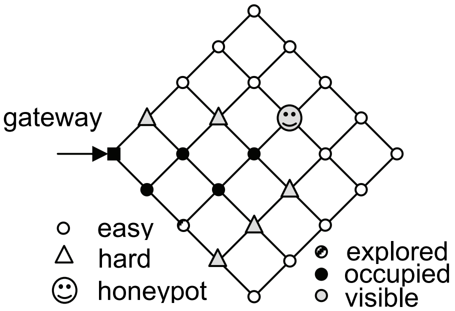

Figure 3. Simulated network intrusion.

Simulated emotional feelings can be used for self-control and guidance of behavior

Most current approaches that enable emotions in artifacts are concerned with behavioral recognition and/or expression of affects [6, 7, 36]. Speculations about adding emotionality as a new cognitive dimension to internal states of the agent itself [13] consider emotional biases as something that potentially can be added, if useful. For example, adding a bias may be justified when a surprise situation in the environment calls for immediate emergency actions. In contrast, we propose that emotional biases are "already there". Indeed, computer behavior typically elicits experience of specific emotional qualia in a human user ("boring", "funny", "dramatic", "exciting", etc.). The question in each case is therefore how to correct for an existing bias rather than how to introduce a qualitatively new bias. The necessary self-correction can be achieved in a cognitive system based on simulated emotional feelings.

In order to do this, the system will learn to simulate emotional feelings in response to its own states. This process requires (i) the notion of a self and (ii) at least two mental perspectives of the Self processed in parallel. In this approach, the biased cognitive process constitutes the content of the mental state *I-Now*. The other, meta-cognitive mental state labeled *I-Feel*, is concerned with monitoring of *I-Now* and recognition of apparent emotional biases in it. Here monitoring means reflecting the content (states) of *I-Now* in the perspective of *I-Feel*, while the recognition of emotional bias is achieved after an associative supervised learning in the emotional cognitive map that receives its input from *I-Now*. The process of correction takes place within *I-Feel*: states copied from *I-Now* are modified or terminated, until the overall configuration of the mental state *I-Feel* is recognized by the cognitive map as neutral. When a critical bias in dynamics of *I-Now* is detected by the cognitive map, *I-Now* and *I-Feel* switch their labels and functional roles. *I-Feel* takes control of the situation, becoming *I-Now*, while the former *I-Now* becomes *I-Feel* to simulate cognition that would take place without self-control, suggesting ideas to *I-Now*. A counterpart process, when an emotional bias needs to be developed rather than eliminated, is described in [33]. Other potentially useful general paradigms involving simulated emotions may include pretend-play (cf. the previous example), understanding other minds in social interactions, robust response to emergency situations, etc. Furthermore, associating emotional values with individual schemas based on experience may help to select appropriate schemas in new situations.

Discussion

First and foremost, we should emphasize that our approach is not competing with the existing state of the art in brain sciences and in artificial intelligence. On the contrary, we are building the upper level on the state of the art and its underlying developments. As we lay the ground for achieving our goal, we bring together recent advances in cognitive psychology, in cognitive and computational neuroscience, and in artificial intelligence. At the same time, there are a number of features that separate our approach from parallel, related approaches, making our approach unique and superior. Here we provide an abridged summary of these features.

Most existing implementations of self-representation and Theory of Mind in artificial intelligence conform to a theory-theory view [11] rather than simulationism [10, 23]. In the hierarchy of modern intelligent agent architectures (reflexive, reactive, proactive or deliberative, reflective, self-aware or meta-reflective), self-aware systems stand on the top; however, existing approaches to their implementation utilize self-concepts limited to notions of self-reference in communications, the distinction between own body and the rest of the world, deictic reference resolution based on formal processing of egocentric perspectives, and simple-minded attribution of goals and beliefs to other entities [24, 34, 8]. These are merely superficial byproducts of the human sense of Self.

We take a simulationist stance [33] in the present work. The novelty of our approach to implementation of a Self based on [33] constitutes the main distinctive feature of our cognitive architecture: (1) the Self is introduced as an imaginary abstraction that does not reflect the actual nature of the cognitive system (e.g. its modular structure and algorithmic rules of dynamics), instead, it represents idealized beliefs about the system's Self, (2) it is implemented in multiple instances that are processed in parallel and interact with each other, while being constrained by the above beliefs taken as axioms, (3) these principles constitute a framework for all forms of higher cognitive activity in our architecture, resulting in additional, emergent properties that are characteristic of a human self and vital for cognitive growth.

The foregoing limited, superficial byproducts of having a human-like Self explored by previous works may well be conceived without any anthropomorphic notion of a Self, and can be exhibited by an artifact without any human-like Self implemented in the system. Examples include: self-reference in communications (reproduced in computational linguistics), the distinction between own body and the rest of the world (materialized in robotics), the attribution of goals, beliefs and intentions to an agent (implemented a while ago in the BDI architecture), analysis of competitive or cooperative behavior in game theory, representations of specific cognitive processes by tokens at a meta-level in various simplistic models of meta-cognition. It is true that any specific feature of this sort taken in a specific problem context can be implemented based on existing approaches; however, a conceptually new, higher-level approach is required for a general-purpose implementation of the entire complex of the key features of the human Self (a) - (e) listed above. Our work offers a solution to this problem.

Addressing the perspectives of implementation of the proposed architecture, we should point to potential technological limitations and their possible solutions. 1°. As rigorous analytical calculations show, neural networks implementing multidimensional manifolds as attractor sets are characterized by a very limited capacity: the maximal number of stored manifolds may be three decimal orders smaller than the number of synaptic connections per neuron [5]. Therefore, a substantial number of neuronal units may be necessary in order to implement a cognitive map isomorphic to a multidimensional manifold: e.g., a space of possible emotions used as an indexing set for schemas and episodic memories. 2°. As a consequence, using this network to recognize symbolic working memory patterns may become another technical problem. A solution to 1° and 2° could be to use supercomputers or specialized neural-network hardware for implementation of cognitive maps and their interfaces with symbolic components. 3°. We can see a potential difficulty with designing a universal standard for the format of representation of schemas that will be used by various implementations of the architecture, regardless of the domain of application. In our view, the right strategy is to provide one universal solution possessing the versatility and sustainability of LISP rather than many case-specific "quick-and-dirty" solutions. 4°. The problem of the parallel cognitive growth and hierarchical self-organization of symbolic and neuromorphic components deserves separate attention here. Using a spatial mapping analogy, the solution can be explicated as follows: arranging buildings on a plane and linking them into districts can be done in parallel by a growing hierarchical cognitive map architecture, as described and studied previously [26]. As a result, connected clusters of 'frozen' mental states may be disconnected from each other in episodic memory. Finding them is another problem to be solved with a contextual cognitive map. On the other hand, a counterpart function implemented in conceptual cognitive maps would allow for modeling *epistemic states* (understood as states of semantic memory, i.e., sets of available schemas) of other agents.

We expect that upon implementation of the proposed self-aware cognitive architecture, a new category of intelligent cognitive systems will emerge with the ability to grow: i.e., to acquire new skills and knowledge autonomously, without an upper limit on the level of acquired intelligence capabilities. This category of systems, during their ontogeny (the period after their initial 'embryo' stage and before their 'maturity'), will be neither programmed, nor evolved in an evolutionary computation style, nor trained by multiple repetitions as an artificial neural network. Instead, they will be taught like humans, using (a) specially designed training facilities, (b) verbal instructions and guidance, (c) behavioral demonstrations, and (d) guidance in developing a personal system of values and skills of self-analysis, meta-learning, reconsolidation of episodic memories, etc. It will be possible to use the new systems in a variety of cognitive paradigms requiring robust learning abilities, communicational and social capabilities in ad hoc teams including humans, and self-awareness. We see as in a near future, an idealized, robust, epigenetic Self in a robot associated with its personal ideals will take over the function of a manually written program, an explicit verbal instruction, or an innate goal.

Acknowledgements. We are grateful to Drs. G. A. Ascoli, L. Nadel, A. B. Butler, H. J. Hamburger, M. D. Bugajska and A. C. Schultz for valuable discussions.

References

[1] Anderson, J.R., and Lebiere, C. 1998. *The Atomic Components of Thought.* Mahwah: Lawrence Erlbaum.

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

[3] Banzhaf, W., Nordin, P., Keller, R.E., and Francone, F.D. 1998. *Genetic Programming ~ An Introduction: On the Automatic Evolution of Computer Programs and Its Applications.* San Francisco, CA: Morgan Kaufmann.

[4] Barresi, G. 2001. Extending self-consciousness into the future. In Moore, C., and Lemmon, K. (Eds.). *The Self in Time: Developmental Perspectives,* pp. 141-161. Mahwah, NJ: Lawrence Erlbaum Associates.

[5] Battaglia, F.P., and Treves, A. 1998. Attractor neural networks storing multiple space representations: A model for hippocampal place fields. *Physical Review* E 58 (6): 7738-7753.

[6] Bosma, H.A., Kunnen, E.S. 2001. *Identity and emotion: Development through self-organization.* Paris: Cambridge University Press.

[7] Breazeal, C. 2003. Emotion and sociable humanoid robots. *International Journal of Human Computer Studies* 59: 119-155.

[8] Cassimatis, N.L., Trafton, J.G., Bugajska, M.D., and Schultz, A.C. 2004 Integrating cognition, perception and action through mental simulation in robots. *Journal of Robotics and Autonomous Systems*, 49 (1-2): 13-23.

[9] Cohen, N.J., Eichenbaum, H. 1993. *Memory, Amnesia and the Hippocampal System.* Cambridge, MA: The MIT Press.

[10] Goldman, A.I. 1993. The psychology of folk psychology. *Behavioral and Brain Sciences* 16: 15-28.

[11] Gopnik, A., and Meltzoff, A. 1997. *Words, Thoughts and Theories.* Cambridge: MIT Press.

[12] Hertz, J., Krough, A., and Palmer, R.G. 1999. *Introduction to the Theory of Neural Computation.* Addison-Wesley.

[13] Hudlicka, E. 2003. To feel or not to feel: The role of affect in human–computer interaction. *International Journal of Human Computer Studies* 59: 1-32.

[14] Koza, J.R. 1998. Genetic Programming II: Automatic Discovery of Reusable Programs. Cambridge, MA: The MIT Press.

[15] Laird, J.E., Rosenbloom, P.S., and Newell, A. 1986. *Universal Subgoaling and Chunking: The Automatic Generation and Leaerning of Goal Hierarchies.* Boston: Kluwer.

[16] Lisman, J.E., and Idiart, M.A.P. 1995. Storage of 7+/-2 short-term memories in oscillatory subcycles. *Science* 267 (5203): 1512-1515.

[17] Meyer, D.E., and Kieras, D.E. 1997. A computational theory of executive cognitive processes and multiple task performance: Part I. Basic mechanisms. *Psychological Review* 63, 81-97.

[18] Miller, G.A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63: 81-97.

[19] Nadel, L., and Moscovitch, M. 1997. Memory consolidation, retrograde amnesia and the hippocampal complex. *Current Opinion in Neurobiology* 7: 217-227.

[20] Nadel, L., and Moscovitch, M. 2001. The hippocampal complex and long-term memory revisited. *Trends in Cognitive Sciences*, 5: 228-230.

[21] Nadel, L., Samsonovich, A., Ryan, L., and Moscovitch, M. 2000. Multiple thace theory of human memory: Computational, neuroimaging, and neuropsychological results. *Hippocampus* 10 (4): 352-368.

[22] Newell, A. 1990 *Unified Theories of Cognition.* Cambridge, MA: Harvard University Press.

[23] Nichols, S., and Stich, S. 2003. *Mindreading: An Intergrated Account of Pretence, Self-Awareness, and Understanding Other Minds.* Oxford: Oxford University Press.

[24] Perlis D. 1997. Consciousness as self-function. *Journal of Consciousness Studies* 4: 509-525.

[25] OKeefe, J., and Nadel, L. 1978. *The Hippocampus as a Cognitive Map.* Clarendon, New York, NY.

[26] Samsonovich, A. 1998. Hierarchical multichart model of the hippocampal spatial map. *Proceedings of the 5th Joint Symposium on Neural Computation*, pp. 140-147. San Diego, CA: UCSD.

[27] Samsonovich, A.V., and Ascoli, G.A. 2005. A simple neural network model of the hippocampus suggesting its pathfinding role in episodic memory retrieval. *Learning & Memory* 12: 193-208.

[28] Samsonovich, A.V., and Ascoli, G.A. 2005 The conscious self: Ontology, epistemology, and the mirror quest. *Cortex* 42 (5), 18 p., in press, published online.

[29] Samsonovich, A.V., and De Jong, K.A. 2003. Meta-cognitive architecture for team agents. In Alterman, R., and Kirsh, D. (Eds.). *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*, pp. 1029-1034. Boston, MA: Cognitive Science Society.

[30] Samsonovich, A.V., and De Jong, K.A. 2004. A general-purpose computational model of the conscious mind. In K. Forbus, D. Gentner, & T. Reigier (Eds.). *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society*, pp. 382-383. Mahwah, NJ: Lawrence Erlbaum.

[31] Samsonovich, A.V., and DeJong, K.A. 2005. Pricing the 'free lunch' of meta-evolution. In: Beyer, H.-G., O'Reilly, U.-M., Arnold, D.V., Banzhaf, W., Blum, C., Bonabeau, E.W., Cantu Paz, E., Dasgupta, D., Deb, K., Foster, J.A., deJong, E.D., Lipson, H., Llora, X., Mancoridis, S., Pelikan, M., Raidl, G.R., Soule, T., Tyrrell, A., Watson, J.-P., and Zitzler, E. (Eds.). *Proceedings of the Genetic and Evolutionary Computation Conference: GECCO-2005*, vol. 2, pp. 1355-1362. ACM Press: New York.

[32] Samsonovich, A., and McNaughton, B.L. 1997. Path integration and cognitive mapping in a continuous attractor neural network model. *Journal of Neuroscience* 17 (15): 5900-5920.

[33] Samsonovich, A.V., and Nadel, L. 2005. Fundamental principles and mechanisms of the conscious self. *Cortex* 42 (5): 669-689.

[34] Shapiro, S.C. 2004. Interests and background relevant to self-aware computer systems, a position statement for the DARPA *Workshop on Self-Aware Computer Systems*, Washington, DC, April 27-28.

[35] Teyler, T.J., and Discenna, P. 1986. The hippocampal memory indexing theory. *Behavioral Neuroscience* 100: 147-154.

[36] Trappl, R., Petta, P., and Payr, S. (Eds.), 2002. *Emotions in Humans and Artifacts.* Cambridge, MA: MIT Press.

[37] Tulving, E. 1983. *Elements of Episodic Memory.* Oxford: Oxford University Press.

[38] Tulving, E. 2002. Episodic memory: From mind to brain. *Annual Reviews in Psychology* 53: 1-25.

[39] Wheeler, M. A., Stuss, D. T., and Tulving, E. 1997. Toward a theory of episodic memory: The frontal lobes and autonoetic consciousness. *Psychological Bulletin* 121(3): 331-354.