Biologically Inspired Cognitive Architecture for Socially Competent Agents

Alexei V. Samsonovich

Krasnow Institute for Advanced Study
George Mason University
Fairfax VA 22030-4444
asamsono@gmu.edu

Abstract

The challenge addressed here is to design a hybrid cognitive architecture that will possess the minimal Core Cognitive Competency (CCC) sufficient for the cognitive and social growth of an agent up to the human level of intelligence, based on autonomous learning. The approach is based on the integration of high-level symbolic (schema-based) and connectionist components at the top representational level, as described previously. The present work specifically addresses the mechanisms of perception, voluntary action and social communication based on the given formalism of schemas. Social competency appears to be vital for CCC.

Introduction

In the near future, machines are expected to make a transition from tools to partners in our society (Albus and Meystel 2001, Samsonovich and Ascoli 2002, Kurzweil 1999, 2005). One scientific aspect of this transition is the following underlying hypothesis, which is not always acknowledged, yet in our view is always implied by any ambitious human-level AI project and is subject to verification by implementation. This hypothesis is: *there is a limited set of cognitive abilities, such that, once implemented in an agent, they will allow this agent to grow cognitively and socially up to the human level of intelligence*, understood as intelligence in general rather than intelligence in a specific domain. This growth will occur through the process of autonomous human-like learning from personal experience, from instructors and partners, from training activities, from public resources of knowledge, etc. This, if implicit, hypothesis allows us to think that the human- level, self-sustainable AI is feasible in the near future, and suggests that a certain set of cognitive abilities, or *Core Cognitive Competency* (CCC), is the key to it. The question #1 therefore is: what are these cognitive abilities, CCC, that enable virtually unlimited cognitive and social growth in an agent, and how one would go about reproducing them in an artifact? In contrast with the Newell's list (Newell 1990), CCC should define the *minimal cognitive embryo* that constitutes a "critical mass" sufficient for the development of human-level intelligence in an artifact. In our view, CCC includes at least the following features.

(a) The following features can be judged based on behavior, using human-like psychometrics (Samsonovich, Ascoli and De Jong, 2006a):
- human-like attention control in working memory,
- understanding (sense-making) in perception of new information and in recall of memories,
- the ability to make independent decisions and to act voluntarily,
- the ability to develop own system of values and generate long-term goals,
- the ability to acquire new conceptual knowledge,
- the ability to remember personal experiences and to learn from them,
- the ability to imagine and to judge possible scenarios and alternative states of mind.

(b) The following features are judged based on the functional organization and implementation of the agent (Samsonovich, Ascoli and De Jong, 2006b):
- the basic human kinds of memory (working, episodic, semantic, procedural),
- the general notion of agency based on a Self-concept (Samsonovich and Nadel 2005),
- Theory of Mind, awareness of others and own Self as agents used in metacognition,
- human-like social intelligence based on internal cognitive modeling of human selves,
- human-like communicational capabilities at the level of semantic input-output,
- human-like emotional intelligence available at the higher semantic level.

The CCC hypothesis says that having all these features available in an agent embedded in a realistic social environment will enable virtually unlimited autonomous cognitive and social growth of the agent. In the present work we argue that the minimal, core social competency is a centerpiece and a vital part of CCC, and we elaborate further our previously proposed cognitive architecture (Samsonovich and De Jong, 2005a) as a potential basis for building socially competent intelligent agents.

Architecture

The approach pursued in this work is based on the integration of symbolic and connectionist components at the top representational level (Samsonovich and De Jong, 2005a). On the symbolic side of this integration scheme, the key elements are a unique, central notion of a *Self* and a formal representation system based on the innovative building block called here a *schema* (Samsonovich and De Jong, 2002, 2003, 2005b). On the connectionist side, the key elements are neuromorphic *cognitive map* components: neural networks that provide for associative indexing of symbolic memories, path-finding in modeled cognitive spaces (Samsonovich and Ascoli 2005a), reinforcement learning, and other functions some of which are discussed below. The integration is achieved through an associative learning mechanism between the symbolic and neuromorphic components.

In the nutshell, the architecture has eight components that are highly interconnected to each other (Figure 1, see Samsonovich and De Jong 2005a): working memory, episodic memory, semantic memory, input-output buffer (all these four components operate using symbolic representations: schemas and their instances, see below), procedural memory (that consists of hard-coded functions, variables and operators collectively called here *primitives*), the driving engine (i.e., an "operating system" that runs all components), the reward and punishment system (used in reinforcement learning and as the origin of goal-directed behavior), and the neuromorphic component implementing cognitive maps.

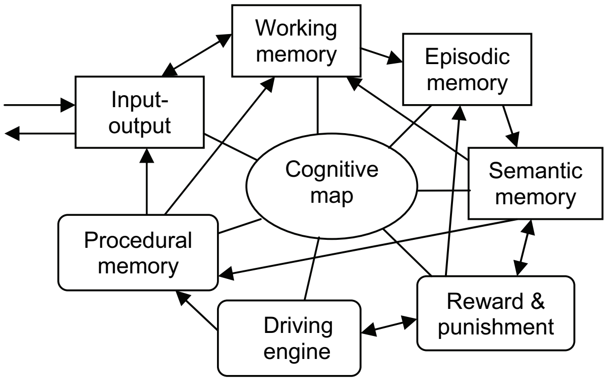

Figure 1 (from Samsonovich and De Jong 2005a). Cognitive architecture at a large scale (see text). Rectangle: higher-level symbolic components, oval: analog connectionist components, rounded rectangle: algorithmic. Arrows indicate essential data flow.

Mental State Lattice and the Self

The notion of a self of an agent has many facets. One popular topic nowadays is the origin of the *sense of self* and *self-awareness* (Damasio, 1999), another side of which is the logic of self-reference (Perlis, 1997). In contrast, here we focus on the *subject-and-author* aspect of the Self (Samsonovich and Ascoli 2005b) that may underlie cognition even in the absence of explicit self-awareness (Samsonovich and Nadel 2005). We understand the Self of a cognitive system as an imaginary unit that is based on a set of fundamental false beliefs about the cognitive system (Samsonovich and Nadel, 2005) called here *self axioms*. This unit-abstraction is represented in the system in order to guide all cognitive information processing at the higher level. The set of self-axioms (semantic constraints imposed on possible representations) can be summarized as follows:

- The Self is a unit that exists as an abstract entity associated with a specific cognitive system. This unit is unique and the same in all circumstances.
- This unit is indivisible and opaque: it has no internal mechanisms, parts or substructure that may be analyzed to account for its behavior.
- This unit is the only subject of all first-hand present and past experiences represented in the system.
- This unit is the only one, independent author of all self-initiated actions of the cognitive system.
- This unit is self-consistent over time in that it relies on its persistent personal values and memories.
- This unit is always localized in one given context (including time, space and agent's body). Therefore, an instance of a Self defines a *mental perspective*.
- This unit is capable of awareness of self and others.

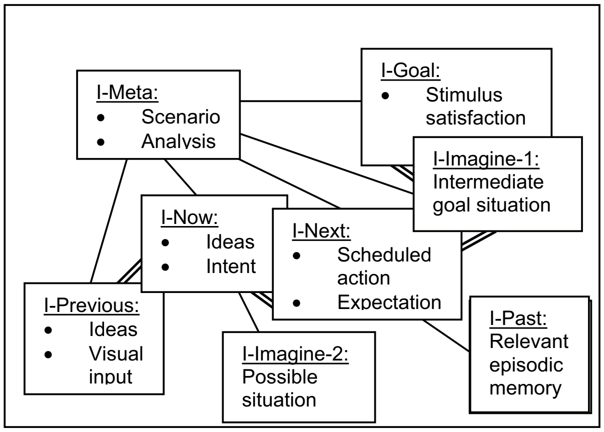

Figure 2. Working memory is partitioned into mental states (boxes), the main sequence of which (double line) makes a working scenario. Each bulleted entry is an instance of a schema, the underlined header in each box is the mental state label.

These axioms apply to what the system represents (beliefs), and not to what the system is. In addition to this static set of beliefs, our notion of Self involves a dynamic (emergent) aspect as well. As illustrated in Figure 2, the Self in our model has discrete instances (e.g., "I" taken at different moments of time) that are represented by simple tokens (labels "I-Now", "I-Previous", etc.).

In essence, implementation of the Self in our architecture is based on the sequence of conceptual steps that introduce several building blocks: (i) self axioms, (ii) a mental state, (iii) the lattice of mental states, (iv) working scenario, (v) working and episodic memory systems, and (vi) cognitive maps that implement personal value systems.

Implementation of self-axioms formulated above begins with the notion of a *mental state*. Working memory is dynamically partitioned into mental states (Figure 2). Each mental state corresponds to an instance of a self and represents the mental perspective of that instance, which is partially captured by a self-explanatory mental state label (e.g., "I-Now", "I-Next", "I-Previous", "I-Imagined", "I-Goal", "I-Meta", "I-Past", "He-Now", "He-Next", "She-Now", and so on). At the same time, the labeling may be ambiguous and is used only to articulate the principles of our framework. Technically, each mental perspective is represented by the associated location in the contextual cognitive map (see below). None of the mental states are permanent architectural components: mental states may emerge and disappear, move across memory systems, change their labels, functional roles and relationships; however, a certain set of standard labels are likely to be found at any given time in an awake system. Among them are I-Now and I-Next: they represent states of awareness of the subject at present and at the next moment.

Active mental states in working memory form a dynamical graph on the lattice of all possible mental perspectives (called the *mental state lattice*: Figure 2). The content of each mental state is a set of instances of schemas bound to the given mental perspective and to each other. The framework allows the system to process each mental state from another mental state (perspective), thereby providing a basis for various forms of meta-cognition. The underlying schema-oriented representation formalism that was described previously at an abstract level (Samsonovich and De Jong 2005a; see below) is specifically designed to enable cognitive growth capabilities and to allow for indexing, retrieval, and organization of memories by the connectionist component.

The subset of active mental states is not randomly selected among all possibilities. E.g., it contains the main sequence of mental states called the *working scenario* that includes I-Now and (in the case of goal-directed behavior) I-Goal. Intuitively, it corresponds to the notion of a stream of consciousness. Mental states in working scenario represent the most realistic and truthful, according to the agent's beliefs, succession of events of the near past, the present and the near future that are relevant to the current perspective of the agent.

Mental states change their perspectives over time: e.g., under normal conditions I-Next becomes I-Now, which then becomes I-Previous, which eventually becomes I-Past. At this stage the mental state is de-activated and moves from working memory to episodic memory. Therefore, episodic memory formation in this architecture could be as simple as deactivation and storage of mental states found in working memory that are about to expire. On the other hand, the process of episodic retrieval is in general more complicated and may require pathfinding (Samsonovich and Ascoli 2005a). In addition to memory of the actual past, episodic memory represents imaginary experiences such as previously considered goal situations. Relationships between mental states in episodic memory are organized based on the same lattice of mental states.

While the volume of working memory may be limited to a reasonably small number of mental states, like the "magic number" seven plus-minus two (Miller 1956, Lisman and Idiart 1995), the volume of episodic memory is virtually unlimited: the system should be able to remember its entire life experience. Semantic memory that stores schemas also has virtually unlimited volume: the system should be able to learn new concepts at any age, while remembering all previously learned concepts.

Mental states in episodic memory can be viewed as organized into short sequences or clusters called *episodes* that once were co-active in working memory. In this sense, retrieval of an episode from episodic memory amounts to retrieval of a past state of working memory (although upon retrieval it is not exactly the same state as it was at the time of experience: e.g., it is perceived as the past). The entire episodic memory, however, may not store one global sequence of mental states: due to many reasons, the sequence will be frequently interrupted. As a result, typical episodes will be stored "as cherries in a bowl". Their proximity on the contextual cognitive map (see below) has more to do with the similarity of their contexts rather than with their proximity in time.

In summary, the Self in this framework is an imaginary abstraction that has multiple representations – tokens to which experiences are attributed. Each of these tokens together with all attributed to it experiences forms a mental state. Therefore, multiple instances of the Self (one for each currently experienced mental perspective) correspond to simultaneously active mental states that are processed in parallel. Representations of the Self and its experiences are constrained by self axioms, producing an illusion that there is an "alive subject" controlling the system.

The question of how the notion of consciousness maps onto this functionalist framework is not of our concern in the present work, yet it deserves to be briefly mentioned. One possibility is to identify the content of I-Now with the content of consciousness. Another possibility is to say that consciousness may include some nearby mental states as well, yet not the entire working memory, but then the criterion for separation remains unclear. A more traditional interpretation of consciousness as working memory (Baddeley 1986, Baars 1988) would not account for higher cognitive phenomena that may happen without consciousness (e.g., Samsonovich 2000).

Cognitive Map

The notion of a cognitive map in cognitive psychology was introduced by O'Keefe and Nadel (1978). A *cognitive map* in the present framework is understood as an abstract metric space that is used as an associative index of symbolic representations, including mental states, schemas and their instances, and can be extended to index elements of schemas. Cognitive maps can be implemented in neural networks, e.g., as quasi-continuous attractors. In this case, points of the map would correspond to patterns of network activity, and Hebbian learning could be used for linking the map to symbolic representations.

The neuromorphic cognitive map component in Figure 1 includes three kinds of cognitive maps: a contextual map that indexes episodic memories, a conceptual map that indexes semantic memories (schemas), and an emotional map that may reflect on values and affective aspects of episodes as well as semantic knowledge. Examples of dimensions would be: time and location for a contextual map; complexity, specificity, rationality, etc. for a conceptual map; valence, arousal and strength for an emotional map.

Cognitive maps can serve many vital functions in our architecture. For example, emotional maps can be used to generate representations of feelings and emotions based on the content of a mental state. Moreover, the valence dimension of the emotional map can be used in reward-punishment mechanisms (e.g., reinforcement learning) and in goal generation during imagery, while the arousal dimension can be used in mechanisms of attention control. Conceptual maps can be used to suggest relevant schemas by associations during cognitive analysis or planning, thereby guiding the process of "thinking". The same mechanisms in combination with other tools can be used in analogy search. A contextual map can be used for pathfinding (Samsonovich and Ascoli 2005a) during episodic memory retrieval, as well as in working memory during planning. More generally, conceptual and contextual maps will help to manage long-term memory consolidation, including generalization, categorization, multiple trace formation, etc.

Again, a unique feature of our approach is that we place cognitive maps and implementing them neural networks on the top of the symbolic part of the architecture; therefore, these neuromorphic components play the integrating role by learning abstract cognitive spaces from symbolic components and then using them as metrics. Together cognitive maps provide an infrastructure for symbolic representations in our architecture. Related to cognitive maps 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. The main structural part of the cognitive map in the brain is the hippocampus.

Schema Formalism

Our notion of a schema introduced previously (Samsonovich and De Jong, 2005a) is a formal system of representation that underlies all our symbolic components. This formalism of schemas can be viewed as a generalized production system that offers both a meta-level of information processing and the ability to incorporate non-symbolic primitives into schemas. 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. E.g., our schemas include concepts, elementary actions and their sequences, as well as templates for instantiations of qualia and abstract notions. The main distinctive feature of our system of schemas is its cognitive growth capability: schemas capturing qualitatively new cognitive abilities can be created autonomously by existing schemas (Figure 3; cf. Laird, Rosenbloom and Newell, 1986). In this sense, schemas are divided into innate (pre-programmed) and acquired (automatically created by the system). As we see, the system of schemas can evolve.

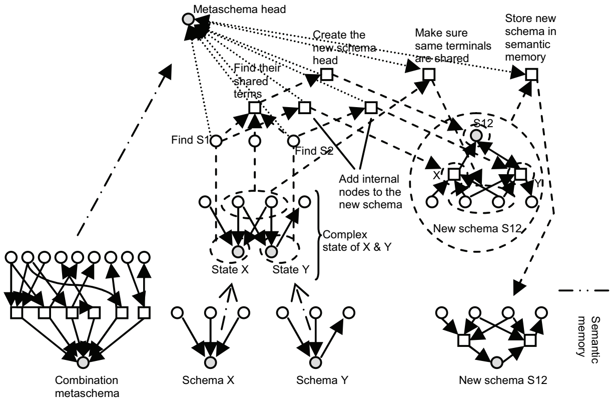

Figure 3. Example of a *metaschema* (i.e., a schema operating on other schemas) used to create new schemas by combining bound instances of existing schemas.

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. Internal nodes, if present in a schema, may represent other schemas and primitives implementing conditions of state creation and effects of state execution. The first step of a state creation based on a schema is called binding. Links of the schema graph together with attributes of nodes specify how bindings of nodes should be established. For example, in Figure 3, schema X is used to create a state in working memory. This state is a copy of X with its terminals bound to some content in working memory, including terminals of state Y.

During the binding of state Y, its last terminal is not used to match a pre-existing content, but is used to create a new content in working memory: these conditions (indicated by directions of arrows in Figure 3) are specified by the attribute “mode of binding”, along with other parameters.

Cognitive Architecture in Action: Numerical Simulation Results

Perception and Understanding

We start our consideration with an example of visual perception of a cube based on three schemas: a vertex, an edge and a face. This example is a convenient starting point for us to develop an understanding of the key concepts of the schema formalism. Intuitively, one can represent a 2-D projection of a wire diagram of a cube as 12 edges. This is the input to the system in our example.

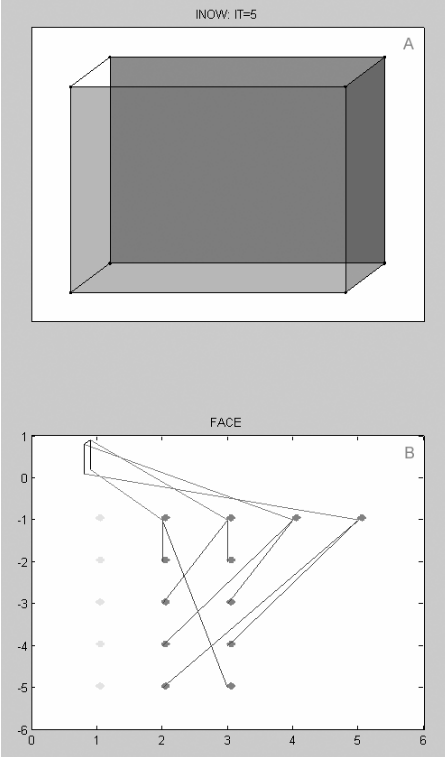

Figure 4. Results of a Matlab implementation of the architecture simulating the perception of a cube. **A:** After five iterations, three instances of the face schema are successfully bound to the available content of awareness in I-Now. **B:** Diagram representing the structure of the face schema together with its internal and external bindings.

The process of *understanding* in our case involves mapping schemas of elements of the cube (e.g., faces: Figure 4 B) onto the wire diagram of the cube that was imported into I-Now from the I/O buffer during the process of *perception*.

Generally, we represent any schema as a two-dimensional array of nodes. All nodes have one and the same, standard set of attributes: their incomplete list is given in Table 1, explanations are given below in the text.

Table 1. Standard attributes of schema nodes.

Attribute	**Nickname**
Attitude	*att*
Perspective	*per*
Stage of processing	*stp*
Category	*cat*
Tag	*tag*
Mode of binding	*mob*
Method of binding	*met*
Bindings	*bnd*
Value	*val*
Quantifier	*qua*
Activation	*act*
Attention	*atn*
Fitness	*fit*

The first (top) row of nodes are terminal nodes, the first of which is the head. The following rows correspond to the top rows of other schemas. Among terminal nodes, there may be those that need to be bound in order for the schema to be instantiated (*stp* = 1) and those that may be bound at a later stage (*stp* > 1). Depending on the *mob* value, a node can serve as an input or an output terminal. Input nodes match some of their attribute values (e.g., *cat*) with those of the nodes of other (pre-existing) instances to which they bind. Following the list, *met* can be deductive, inductive or abductive; *bnd* is a pointer to another node; *val* may contain a numerical or symbolic value associated with the category; *qua*, if specified, allows a node to be bound to a set of nodes; *act* and *atn* determine the probability of selection of the node for processing (the processing of schemas is done based on stochastic rules in parallel); *tag* is used to label all instances of one and the same object (to distinguish them from other objects of the same category); *per* is the mental perspective (an instance of a Self) to which the instance of the schema is attributed; *att* is the position of the instance with respect to the subject in the cognitive space; *fit* is used in the process of evolution of the set of schemas. *Semantic memory* in this framework, in addition to the entire set of schemas, includes the hierarchy (more generally, a semantic network) of all categories represented by schemas and a *reference memory* that is based on a spatial map of the environment and represents the current (believed by the agent) configuration of the world. Furthermore, semantic memory can be partitioned into domains of knowledge and is associated with conceptual and emotional cognitive maps.

These general principles of our framework apply to all embodiments of the architecture, including the current example of perception of a cube. E.g., the top row of nodes of the face schema (Figure 4 B) have the following categories, from left to right: “face”, “vertex”, “vertex”, “vertex”, “vertex”. Each of the following rows in Figure 4 B is the schema of an edge, with categories of nodes, from left to right: “edge”, “vertex”, “vertex”. Figure 4 B shows an instance of the face schema successfully bound to a face of the cube.

A next step in this process of “understanding” of the cube would be to assign the depth (the third coordinate) to tilted faces. The latter operation is ambiguous. Therefore, as the time flows and the mental perspectives change, flipping from one perceived configuration of the cube to another occurs spontaneously, and can be pre-determined by voluntary attention control. In one scenario, two mental states take the position of I-Now in turn during this process. This same sort of flipping is observed experimentally in subjects perceiving the Necker cube (Slotnik and Yantis, 2005).

Voluntary Action

Most self-initiated actions produced by our architecture are voluntary actions. Each voluntary action involves a standard set of key elements listed below:

1) generation of *ideas* – feasible actions;
2) preliminary evaluation of the ideas based on their consistency with respect to the current working scenario (e.g., by using a heuristic schema associated with the current goal or desire);
3) elaboration of expectations for selected ideas (this is done in separate mental states I-Imagine);
4) selection of an *intent* (the corresponding I-Imagined acquires the status of I-Next);
5) motivation of the intent (finding one good reason why this should be done: here again the same heuristic can be used);
6) scheduling execution;
7) controlled execution (I-Next becomes I-Now);
8) perception of the outcome and its comparison against the previous expectations (which results in a positive feeling in the case of success);
9) if there is a mismatch, a process of conflict resolution starts in a metacognitive perspective (I-Meta becomes I-Now). This last step could result in revision of previous beliefs of the agent, etc.

To illustrate voluntary actions, here we consider another example of embedding, in which our agent is a car in a virtual city (Figure 5). Space and time are discrete. The agent moves on a rectangular lattice of 20 nodes – intersections. Perceived features of the environment at each moment of time include features of the current intersection only. Available actions are one-step moves in the four directions. E.g., a mental state I-Now of the agent located at the intersection C3 will include the following perceived instances: gas station, pond, supermarket, building. The attitudes of these instances are equal to nil, which means “current, actual fact”. In addition, the mental state may include the current intent, the previous action, imaginary actions, etc., all with appropriate attitudes.

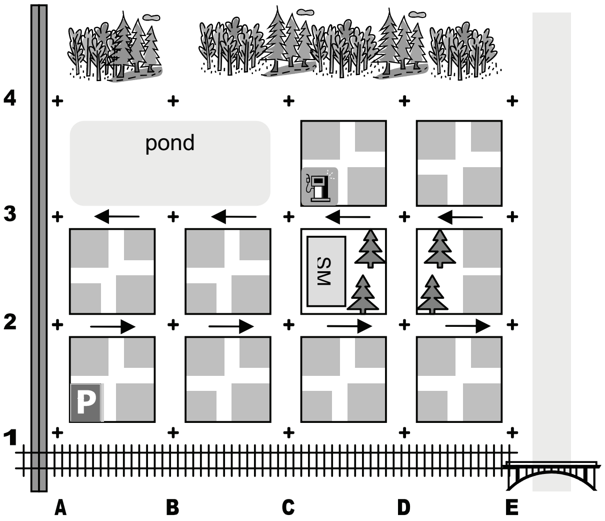

Figure 5. Virtual city map. The city is surrounded by a forest from the north, a highway from the west, a river from the east, and a railroad from the south. The agent moves on the lattice of intersections (crosses), passing one block at a time. Some streets allow for one-way traffic only (arrows). SM: supermarket, P: home parking garage.

Suppose that initially the agent is wondering in the city driven by the desire to explore it: there is no explicit goal, but the idea of seeing new locations dominates in the working scenario (plus the rule of not reversing the steps immediately applies). Therefore, at the step (4) in the list above the preference will be given to ideas of moves that immediately lead to previously unexplored locations, if there are such moves; otherwise the choice will be random. This is a rather trivial example of voluntary behavior. One numerically generated example of an exploratory trajectory is the following:

A1 B1 B2 B3 A3
A2 A1 B1 C1 C2
D2 E2 E3 D3 D4
C4 B4 A4 A3 A2
A1 B1 B2 B3 A3
A4 B4 C4 D4 E4 (1)
E3 E2 E1 D1 C1
B1 B2 B3 A3 A4
B4 C4 C3 B3 A3
A4 B4 C4 C3 C2
C1 D1 D2 D3 C3

By now the agent has explored all locations in the city, most of them multiple times, and associated objects seen and mental states experienced at those locations with the two-dimensional lattice, which is a trivial example of a contextual cognitive map. At the same time, we assume that objects seen together were associated with each other on a conceptual cognitive map. E.g., a gas station (the concept) is associated with a supermarket, a pond and a building. A pond is associated with a gas station, a highway, a forest and a building, and so on.

Suppose that the agent is located at its home parking garage (Figure 5: P) and wants to get to the gas station. We consider three methods of intent generation given the current goal. Planning capabilities will not be assumed (however, they could be learned and represented by schemas during further exploration of the city).

Method A: intent selection using the spatial (contextual) map and a heuristic of minimizing the distance to the goal location. This method in the given case trivially results in a path of 4 steps, which is a shortest possible path.

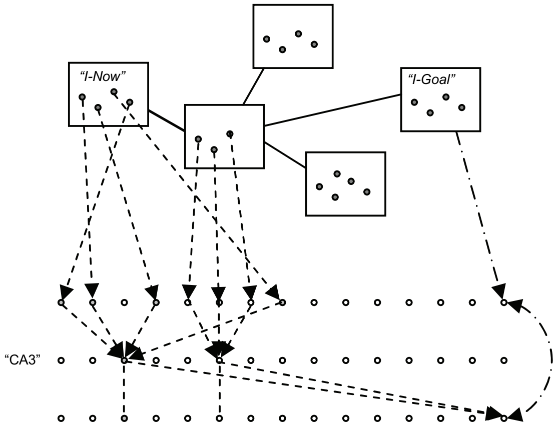

Figure 6. Pathfinding in working memory using a contextual cognitive map Each box represents a mental state, each dot inside a box is an instance of a schema. The neural network component consists of three layers: the interface layer (top of the three, used to associate each mental state with a unique node in each of the other two neural network layers) and the "CA3" and "CA1" layers of the contextual cognitive map (Samsonovich and Ascoli 2005). Connections between mental states represent feasible actions.

Method B: intent selection using associations stored in the conceptual cognitive map. In this case the agent will give preference to moves that lead to locations where more features have (stronger) associations with the features of the goal location. Therefore, the first step will be from A1 to A2 (the highway is associated with the pond which is associated with the gas station, while the railway is not associated with any object that is directly associated with the gas station). The trajectory consists of 6 steps:

A1 A2 A3 A4 B4 C4 C3.

Method C: intent selection using the path-finding mechanism based on a contextual cognitive map. In this case, we assume that our neuromorphic component has a neural network architecture of Figure 6 implementing the pathfinding learning and search algorithm described in detail previously (Samsonovich and Ascoli, 2005a). The result in this case is easy to derive from (1): it is one of the shortest trajectories, A1 B1 C1 C2 C3. Note, however, that if the exploration trajectory (1) were cut in half, then the path found by the pathfinding algorithm would be longer than the path found by the conceptual associative network:

A1 B1 B2 B3 A3 A4 B4 C4 C3.

It is also interesting that all three methods work significantly better than a random search. Nevertheless, these examples are not intended here to demonstrate cognitive and learning abilities of the architecture in simple spatial tasks. Rather, they are viewed as an intuitive illustration of principles underlying our architecture that will work in more complicated, non-spatial learning paradigms. We consider these principles and mechanisms as the key building blocks of our integrated architecture.

Constructing Systems of Values

It is important to understand that principles considered in the previous section are not limited to spatial reasoning. For example, the same or similar mechanisms could be used based on emotional cognitive maps that would arrange concepts and mental states in an abstract space of affects and feelings that may have little in common with Figure 5. On the other hand, a key question is: how to allocate representations in this space?

Despite the long history of study of cognitive maps, their dimensions, intrinsic metrics and functional properties are poorly understood. Given that direct access (e.g. Quiroga et al., 2005) to brain representations of concepts is difficult due to ethical and technological reasons, alternative approaches deserves consideration. As a possible answer to the above question, we consider an example based on linguistic material.

The material used in this part of the study is the thesaurus of English synonyms and antonyms available as a part of Microsoft Office Word 2003. Of the entire dictionary, only the giant connected cluster of 8,000 concepts (including words and short phrases) was selected. The cluster has a structure close to a scale-free network. An average word in it has 1.8 synonyms (3.0 including references to this word) and 0.8 (1.4) antonyms. Distances between words, however, do not correlate well with semantic differences or similarities and could be misleading.

The method of study was the following. Each concept was represented by an abstract particle randomly allocated in a compact multidimensional manifold. Forces of attraction between synonyms and repulsion between antonyms were introduced. Thus defined dynamical system underwent simulated annealing and dimensional reduction procedures. The resultant self-organized configuration was subject to principal component analysis. Finally, the semantics of principal components were identified. Words were sorted along principal components, the very top of the list was compared to the very bottom. Results are represented in Table 2.

Table 2. Sorted List of Words Extracted from the Microsoft Word 2003 Thesaurus.

Sorted by the first principal component, identified as “valence”		Sorted by the second principal component, identified as “arousal” (taken from a different simulation)	
Top of the list	*Bottom pf the list*	*Top of the list*	*Bottom of the list*
found ethics prove decisiveness unquestionable repeat put up excitedly definite ethical come to attraction conclusively have knowledge of treasure cherish elegant admit rejoice resolution aristocratic acclimatize consciousness	sap pointless enfeeble weaken grow weaker inept unenthusiastically useless futile unfit incompetent descent take out hesitation droplet drop withdraw immoral dishearten dull-witted unconsciousness tear down morally wrong fallow	outbreak eruption outburst thump clunk punch alarm clock knock tirade pound thud affect impact rap bash throb fret make an effort pounding taunt stilted concerned	placid simplicity consecutive successive neutrality impartiality calm down untouched unaffected peace easy painless unconcerned calming tranquillity soothe effortless soothing calmness reassurance untroubled composure serenity equanimity alleviate carefree

It is found that a stable, self-organized distribution of concepts (words) in an abstract space obtained with this paradigm can form a cognitive map, in that spatial coordinates of concepts reflect their semantics. The distribution is polarized into two broad, fuzzy clusters of nearly equal size, separated along the main principal component that captures the notion of valenve: “positive vs. negative”. Otherwise, the map is a quasi-continuum, in which at various scales relative locations of concepts correlate with their semantic differences. Interestingly, the strongest two principal components roughly correspond to Osgood’s (1957, 1969) dimensions of the “affective space”: evaluation, potency, and activity.

The main conclusion in this section is that spatial-dynamic analysis of natural language may provide a key to understanding the human system of values and semantic memory, and that similar results can be expected for episodic memories, concepts, or any representation system where synonym and antonym relationships can be defined.

Social Communications

So far we were introducing and studying elements of the architecture step by step, considering only one agent. Now we turn our attention to the most interesting aspect of the presented framework, at which all components should come into play: social interactions among agents. From now on we will be considering two or more architectures of the above kind that interact with each other.

Understanding the dimensionality of the problem

Most of modern natural language processing (NLP) research is concerned with either extracting semantics from a verbal message (natural language understanding: essentially, the disambiguation problem) or representing given semantics by an utterance (natural language generation). We are not going to discuss NLP (Jurafsky and Martin 2000) and related problems here. Instead, we are going to focus on a bigger question: what to do after the semantics of a perceived message are available? Therefore, we assume that problems of extracting semantics from communications, as well as expressing mental states using language, are solved or finessed (Ortiz, 1999; Dragoni et al., 2002).

Following many previous attempts of formalization of this problem (Allen 1995 chapter 17, Panzarasa et al. 2002), here we outline our original approach. For example, suppose that the message was a simple question, like “What time is it?” Suppose that it was established that the true semantics of this message (given the context) is nothing but an inquiry about the current time. What an intelligent social agent should do in response to this message? Should it immediately tell the time? Why? If the agent is telling the time in response to every request like this automatically, without any motivation, analysis or voluntary control, then it seems like there is no big difference between this agent and a talking watch that tells you time whenever you press a button on it. On the other hand, if the agent remains silent, then it could mean that the agent does not understand the question. Where is that element that characterizes intelligence in this situation? No matter how smart the agent is in other specific domains (can beat Kasparov in chess, capable of solving superstring equations), you would not call it socially competent if it does not know what to do when asked a simple question like “What time is it?”

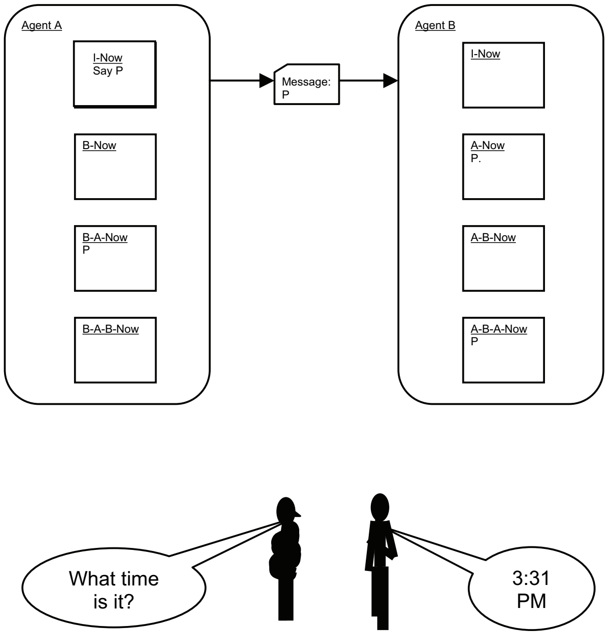

Figure 7. Message passing during social communications may involve higher orders of Theory of Mind – even in simplest cases.

Let us give it a thought. What does it mean that a question, a statement, or any utterance for that matter, occurs as a part of social interaction between A and B, from the point of view of the framework that we have outlined above? Let us assume here that the true semantics of the utterance are available. In general, let us think of messages at the level of semantics, and not below that level. Now, we want to be able to describe social communications between A and B in terms of messages passed from A to B and from B to A, where each message sent from A to B is understood as an event caused by A that consists in the instantiation of certain semantics in the mind of B. More precisely, in terms of our formalism, we would like to think of a certain instance of a schema with a certain mental attitude in the mind of B (in addition, of course, it is natural to think that this process is reflected by awareness in the mind of A, etc.).

For example, suppose that A, who is communicating to B, makes a statement of a fact P, e.g., “You are fat!”. Then B may represent P as an attitude of current awareness attributed to the current mental perspective of A (A-Now in the right box in Figure 7). This scenario implies that B trusts A, in this case at least. But then the story is still not complete. Not only B is aware that A is aware of P, but B is aware that A initiated a message to B about this instance of awareness. It probably should not be a mistake to consider each message passing event during social communication as a voluntary action. Therefore, B is aware that A has committed a speech act, which involves all the key elements of a voluntary action, e.g.: selection of an intent among feasible ideas, understanding the motivation, generation of expectations, evaluation of the expected impact from the point of view of the current working scenario, etc. (see above). As a minimal parsimonious interpretation, B may think that A was motivated by the desire to let B know that P, so that B could use this knowledge. This also implies an assumption that A thought that B was unaware of P, and so forth, not to mention the higher order Theory of Mind (B believes that A believes that B believes that A believes…). Now we see an exponentially growing complexity in multiple dimensions, even with a very simple example.

The situation is similar when a statement of a fact P is replaced by a question Q (e.g., “What time is it?”) or a directive D (e.g., “Give me your valet!”). For example, if the semantics of D is an immediate action of B, then why should B perform (or not perform) the requested action? The answer could be given based on the standard voluntary action schema, if D would be an idea that came to the mind of B independently. On the contrary, we are talking about a situation when D has the definite source: A, and B is aware that not only A is the source of the idea of D, but in addition A has voluntarily communicated this idea to B. Therefore, on the one hand, an imaginary episode of B performing D apparently fits into the current (goal-directed?) working scenario of A, and, moreover, A did not hesitate to tell this directive explicitly to B, while being aware that B may also think of D (independently of A) and has his or her own power to decide what to do. Now, B may consider at least three options: (1) to do D, (2) not to do D, or (3) to send a new message to A in response, before deciding on D. Suppose B decides (voluntarily) to do D. Should this be considered as a favor to A? Even in a situation when both A and B would benefit from B performing D (e.g., imagine that spouses A and B stopped in their car at a banking machine), should nevertheless D be considered a favor to A because A is pretending to be in charge, and B is pretending to obey? And so on. Note that there would be no such questions if D was injected by A into the mind of B using a masked priming technique or hypnosis. Social communication is different from other kinds of communication and involves a concept of a Self as a minimum of social competency.

This situation that emerges even with a simplest directive D passed as a message in social communication is similar to a situation with a simplest question Q, e.g., “What time is it?” or with a simplest observation of a fact P, e.g. “You are fat!”. In conclusion, the resultant global picture appears too complicated even with simplest messages, if all implications need to be addressed. Then, what about messages referring to other mental perspectives, e.g.: “Remember what you told me yesterday?” Of course, it would be a mistake to think that all possible implications necessarily get elaborated and represented in the mind of each agent during social interaction. As with the limited span of awareness in general, representations of social attitudes may be limited by some "magic numbers". Within the limit, some key elements must be present in most cases. What are these key elements? We consider this question next.

Reducing the problem to a simple schema

Among the key elements that are likely (if not necessary) to be involved in any act of message passing during a social communication are the semantics of the message itself, the source agent and the recipient. We probably can agree that the barebone semantics of a message are well defined, if the message is represented by an instance of a schema that is available to both agents (yet the part of the semantics that place an instance of the schema into the right context cannot be dismissed: e.g., a question like "What time is it?" could be a matter of life and death, if A and B are strapped with explosives and have their hands on detonators). The next key element is the target mental perspective in which the semantics will be instantiated by an instance of the schema. For a simple message like "What time is it?" or "Give me your valet" this target perspective will be the current perspective of the source: He-Now or She-Now. As mentioned previously, other mental perspectives may be involved as well, in some cases with a necessity: this will be addressed below. Primarily though, an instance of a schema like "T the current time" (with a variable T in place of the value) will be attributed to the target She-Now (or He-Now). The next key element is the mental attitude of this instance. E.g., in the example "What time is it?" the attitude will be "desire-to-know". In the case "You are fat!" represented as fat→you (i.e., an instance of the property "fat" bound to "you" that needs to be replaced with B) the attitude will be nil, i.e., "current, actual fact".

One element is still missing that has to do with the fact that a message passing event took place, and A voluntarily initiated the message. Messages in real life are not passed telepathically, although the above story as presented so far seems to be consistent with the idea of telepathy. On the contrary, it takes a behavioral motor action to say a sentence or to produce a written word, and this action almost under all circumstances is performed voluntarily. Of course, there is no need to represent an act of verbal communication as motor activity involving lips, tongue, etc. in a simple implementation of the architecture. Instead, an act of message passing may be described as an atomic event that results in the instantiation of the key mental attitude(s), and yet the fact that it is a voluntary action initiated by the source agent needs to be represented explicitly.

Therefore, a schema of an act of message passing in social (not necessarily limited to verbal) communication must belong to the category of voluntary acts and may look as represented in Figure 8. This schema can be used for reasoning about (and production of) own utterances as well as for reasoning about messages received from partners. It contains the minimal key elements that are sufficient for elaboration (when necessary) of the whole complexity of mental representations associated with a dialogue (as indicated above), while at the same time it provides a means of limiting the number of representations without loss of the essential semantics.

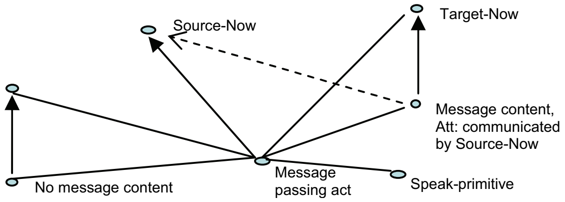

Figure 8. Schema of message passing in social communications. The primary effect of a message is the instantiation of the message content in the Target perspective in the recipient's mind. The schema represents message passing as an event of the transition from a configuration {Target-previous, No message content} to the configuration {Target-now, Message content}: an event that is a voluntary action performed by the Source agent. Source and Target are references to mental states (instances of Self) to which the message content and the act of message passing are attributed. Complex messages may involve multiple target perspectives.

As with any schema processing in our framework, the processing of the message passing schema (Figure 8) is done within one and the same mental state: all Theory-of-Mind aspects in this case are captured by the attitudes of nodes. These attitudes allow the system to copy results into the appropriate mental states (creating them, if necessary) and to continue the processing of information there, should this become necessary for any reason. The bottom line is that with this schema (Figure 8), all the complexity of human-level message handling in communications can be reduced to a simple template. This central idea will be elaborated further elsewhere.

Concluding Remarks

In this work we have elaborated new details of the formalism of schemas that underlies our concept of a biologically-inspired cognitive architecture (Samsonovich and De Jong, 2005a) and presented examples of schemas, explaining how they solve specific cognitive problems in a virtual environment. The architecture presented here offers synergistic integration of symbolic and connectionist approaches. Because of the lack of space, we did not review CLARION (Sun 2004) or ACT-R (Anderson and Lebiere 1998), both of which also claim to offer similar integrations. It could be helpful to the reader to point that our architecture, in contrast with others, offers integration at the top representational level, placing neuromorphic components at the center of the symbolic core. In our case, "top representational level" means that both symbolic and connectionist style representations are co-equals, unlike, eg, ACT-R and CLARION, where the sub-symbolic representations are lower-level than the symbolic ones.

We proposed a hypothesis that there is a limited set of core cognitive abilities, such that, once implemented in an agent, they will allow this agent to grow cognitively and socially up to the human level of intelligence. We outlined our intuitive understanding of these key cognitive abilities and developed further an architecture that allows for their implementation.

We illustrated by numerical examples how the proposed architecture will work in perception and understanding of sensory input, in conceiving and execution of voluntary actions, and in social interactions.

We believe that social growth and therefore, according to our hypothesis stated in the Introduction, the core social competency, is the key to designing intelligent agents that are capable of autonomous cognitive growth up to a human level. While the task of defining what this core competency is may turn even more difficult than the implementation itself, one could effectively "solve" it by defining instead a set of cognitive tests and related metric criteria, such that passing them would be relatively easy for an average human subject and at the same time highly unlikely for a computational agent that does not possess the Core Cognitive Competency, regardless of its interface with the world. These tests and metrics could then successfully guide our future design of cognitive architectures and computational agents. On the other hand, there are examples of cognitive tests and metrics (e.g., the Turing test and its descendants: e.g., Korukonda, 2003) that have not resulted in a substantial advance of the state of the art in AI. Therefore, in order to design the tests, one needs to know specifically (intuitively rather than mathematically) what critical cognitive dimensions need to be captured.

Therefore, in this work we gave a list of the key cognitive dimensions (sometimes called the "magic" of human cognition: i.e., the most general, higher human cognitive abilities that computers still cannot reproduce) that, in our view, a computational agent must have as an individual, independent of any social or environmental context, in order to be able to grow up.

One problem with designing appropriate tests for cognitive architectures that specifically address the critical cognitive dimensions is that during the tests the agent must be able to show its true level of cognitive competency in a given environment, while its vision, NLP or motion skills should not become the bottleneck in the evaluation. How this condition can be achieved? We see one possibility: the appropriate semantics need to be injected into the agent. Based on this idea of finessing the input-output capabilities at an early stage of cognitive architecture design, we envision a successful scenario of incremental growth of an agent, from the cognitive core up to a sophisticated, complete end-to-end architecture, with the gradual addition of the input-output capabilities accompanied by a gradual increase of the realism and relaxation of restrictions in the virtual environment.

Again, it is important to make sure that the agent has the minimal Core Cognitive Competency at the beginning of the process of development rather than to expect its emergence by the end of the growth process that is contingent on it. Accordingly, below we propose a sample set of cognitive tests and metrics that capture some of the 'magic' of human cognition in a way that does not depend on human-level commonsense knowledge, human-level sensory-motor input-output or natural language capabilities. The proposed tests serve an illustrative purpose.

Test A. Assess the ability to simulate multiple mental perspectives, mental attitudes and states of mind, to infer relationships among agents based on their behavior, to understand intentions and goals of individual agents and a team based on their behavior. At this stage, events and objects will be entered into the architecture in a symbolic form, based on available innate concepts (schemas), and processed as atomic units.

Test B. Assess the higher-level social and communicational intelligence based on a specially designed language that will be used for communications with the architecture. In a simplest version, the language may consist of a limited number of enumerated sentences with a priori translated semantics (i.e., translated into the corresponding internal representations of concepts in the architecture). In a more complicated version, the language may consist of a limited vocabulary and a limited set of grammar rules that will not lead to an ambiguity during semantic interpretation of any grammatically correct sentence, so that the language-to-semantic conversion can be easily algorithmized.

Test C. Assess the cognitive growth ability, in particular, the ability to learn new concepts (schemas) based on generalization of limited experience and hypotheses testing, using an exploratory paradigm. The test will involve deliberate actions, e.g., controlled motion in an environment; however, the ability to control self-motion should not be a bottleneck. Therefore, the architecture will operate with macroscopic behavioral building blocks treated as atomic units, assuming their correct execution at the lower level.

We would like to conclude with the following general remark. Building intelligent systems that exhibit the "magic" of human cognition is a difficult, but probably the most interesting and critical challenge of our time. There are many possible views of this challenge, and there are many suggested approaches to its solution. In this paper one particular approach is presented that takes us one step closer to a robust, integrated computational cognitive framework.

Acknowledgments

The author is grateful to Dr. Kenneth A. De Jong, Dr. Giorgio A. Ascoli, Mr. Mark Coletti and Mr. Robert Lakatos for valuable discussions. This work is supported by the DARPA IPTO BICA Grant "An Integrated Self-Aware Cognitive Architecture".

References

Albus, J. S., and Meystel, A. M. 2001. *Engineering of Mind: An Introduction to the Science of Intelligent Systems.* New York: Wiley.

Allen, J. *Natural Language Understanding: Second Edition.* Redwood City, CA: Benjamin/Cummings.

Anderson, J.R., and Lebiere, C. 1998. *The Atomic Components of Thought.* Mahwah, NJ: Lawrence Erlbaum Associates.

Baars, B. J. 1988. *A Cognitive Theory of Consciousness.* Cambridge, MA: Cambridge UP.

Baddeley, A. D. 1986. *Working memory.* Oxford: Oxford University Press.

Damasio, A. R. 1999. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness.* New York: Harcourt.

Dragoni, A.F., Giorgini, P., and Serafini, L. 2002. Mental states recognition from communication. *Journal of Logic and Computation*, 12 (1): 119-136.

Jurafsky, D., and Martin, J. H. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Upper Saddle River, NJ: Prentice Hall.

Korukonda, A.R. 2002. Taking stock of Turing test: a review, analysis, and appraisal of issues surrounding thinking machines. *International Journal of Human-Computer Studies*, 58: 240-257.

Kurzweil, R. 1999. *The Age of Spiritual Machines.* New York: Penguin.

Kurzweil, R. 2005. *The Singularity Is Near: When Humans Transcend Biology.* New York: Penguin.

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

Lisman, J. E., and Idiart, M. A. P. 1995. Storage of 7±2 short-term memories in oscillatory subcycles. *Science* 267: 1512-1515.

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.

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

O'Keefe, J., and Nadel, L. 1978. *The Hippocampus as a Cognitive Map.* New York: Clarendon.

Ortiz, C.L. 1999. Introspective and elaborative processes in rational agents. *Annals of Mathematics and Artificial Intelligence*, 25: 1-34.

Panzarasa, P., Jennings, N.R., and Norman, T.J. 2002. Formalizing collaborative decision making and practical reasoning in multi-agent systems. *Journal of Logic and Computation* 12 (1): 55-117.

Perlis, D. 1997. Consciousness as self-function. *Journal of Consciousness Studies* 4(5/6):509-525.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., and Fried, I. 2005. *Nature* 435(7045):1102-1107.

Samsonovich, A. 2000. Masked-priming 'Sally-Anne' test supports a simulationist view of human theory of mind. In Mel, B.W., & Sejnowski, T., eds. *Proceedings of the 7th Joint Symposium on Neural Computation* (vol. 10, pp. 104-111). San Diego, CA: Institute for Neural Computation, UCSD.

Samsonovich, A., and Ascoli, G. A. 2002. Towards virtual brains. In Ascoli, G. A. ed. *Computational Neuroanatomy: Principles and Methods,* 423-434. Totowa, NJ: Humana.

Samsonovich, A. V., and Ascoli, G. A. 2005a. A simple neural network model of the hippocampus suggesting its pathfinding role in episodic memory retrieval, *Learning and Memory*, vol. 12, pp. 193-208.

Samsonovich, A. V., and Ascoli, G. A. 2005b. The conscious self: Ontology, epistemology and the mirror quest. *Cortex* 41(5):621-636.

Samsonovich, A. V., Ascoli, G. A., and De Jong, K. A. 2006a. Human-level psychometrics for cognitive architectures. In *Proceedings of the International Conference on Development and Learning, ICDL5.* Bloomington, IN. Forthcoming.

Samsonovich, A. V., Ascoli, G. A., and De Jong, K. A. 2006b. Computational assessment of the 'magic' of human cognition. In *Proceedings of the International Joint Conference on Neural Networks IJCNN-2006*. Vancouver, BC. Forthcoming.

Samsonovich, A. V., and De Jong, K. A. 2002. General-purpose meta-cognitive systems: From philosophical ideas to a computational framework. *Artificial Intelligence - Ukraine* 2002(4):67-73.

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 (CogSci2003) 1029-1034. Boston, MA: Cognitive Science Society.

Samsonovich, A. V., and De Jong, K. A., 2005a. Designing a self-aware neuromorphic hybrid. *AAAI-05 Workshop on Modular Construction of Human-Like Intelligence: AAAI Technical Report*, vol. WS-05-08, K. R. Thorisson, H. Vilhjalmsson, and S. Marsela, Eds. Menlo Park, CA: AAAI Press, pp. 71-78.

Samsonovich, A. V., and De Jong, K. A., 2005b. A general-purpose computational model of the conscious mind. In *Proceedings of the Sixth International Conference on Cognitive Modeling*, M. Lovett, C. Schunn, C. Lebiere, and P. Munro, Eds. Mahwah, NJ: Erlbaum, pp. 382-383.

Samsonovich, A. V., and Nadel, L. 2005. Fundamental principles and mechanisms of the conscious self. *Cortex*, vol. 41, pp. 669-689, 2005.

Slotnik, S.D., and Yantis, S. 2005. Common neural substrates for the control and effects of visual attention and perceptual bistability. *Cognitive Brain Research*, 24: 98-108.

Sun, R. 2004. The CLARION cognitive architecture: Extending cognitive modeling to social simulation. In: Ron Sun (Ed.), *Cognition and Multi-Agent Interaction.* Cambridge University Press: New York.