Universal Learner as an Embryo of Computational Consciousness

Alexei V. Samsonovich

Krasnow Institute for Advanced Study, George Mason University
4400 University Drive MS 2A1, Fairfax, VA 22030-4444
asamsono@gmu.edu

Abstract

A universal learner considered here is a cognitive agent that can learn arbitrary schemas (understood broadly as concepts, rules, categories, patterns, values, etc.) as well as associated values, experiences and linguistic primitives, with the help of a human instructor. The agent starts with a limited number of innate schemas and limited innate natural language capabilities. Through the process of bootstrapped cognitive growth, the agent is expected to reach a human level of intelligence in a given domain, plus have the ability to communicate its experiences using natural language. With certain constraints on the functional organization of the agent architecture, this phenomenon can be regarded as an emergent computational consciousness and should be studied as a psychological phenomenon.

Introduction: Universal Learner Is the Key

Computational consciousness can be defined as a fully functional computer-based implementation of features and principles that constitute the essence of human consciousness as we know it subjectively [1]: including the self and its basic mental states, awareness of self and understanding of other minds, mechanisms of voluntary action, the four kinds of memory (working, semantic, episodic and procedural), commonsense knowledge and the general ability to learn, a system of values and feelings and self-consistency over time, all integrated in one embodied cognitive agent architecture.

A *universal learner* considered here is a cognitive agent (an electronic student) that can acquire arbitrary new knowledge from its own experience with the help of an instructor. The agent should be able to use acquired knowledge in further learning, with no a priori limitations on its bootstrapped cognitive growth abilities. In addition, it is assumed that the process of thinking and learning in the agent is constrained by the architecture design in order to make it consistent with human thinking and learning, understood in the functionalist sense.

Trajectory connecting the two notions – an embryonic universal learner and a true computational consciousness – may be long, starting at a relatively small and primitive cognitive system and ending at a very large and powerful one, yet it should not require a programmer intervention in the middle. Given the structure of human knowledge, the above assumptions lead to a prediction that the agent should be able to reach a human level of intelligence virtually in any domain of human expertise that it will be able to explore autonomously, guided by an instructor. This conclusion has strong implications: it means that a breakthrough in artificial intelligence (AI) is possible based on this approach, as soon as an efficient universal learner can be implemented. The purpose of the present work is to explain how a universal learner can be created and to discuss scientific and philosophical implications of emergent computational consciousness. Some necessary conditions for a universal learner are addressed immediately below.

Internal Representations

In order to be able to learn a set of concepts from an instructor, the agent should be able to represent them internally. A problem is to find the right representation system that would allow the agent to store any human knowledge in a practically useful form. We previously described a cognitive architecture [2] inspired by studies of the human brain-mind, which can be considered as a solution to this problem. It has four memory systems found in the human brain: episodic, semantic, working and procedural (Figure 1). They are unified by the cognitive map that indexes memories. Our key building block is a schema that is used to represent concepts in one universal format. It can be conceived as a graph, in which nodes represent cognitive categories. Instances of schemas populate mental states, and mental states populate working and episodic memory systems. As a result, this architecture called GMU BICA can store arbitrary concepts, values and personal experiences in its long-term memory [2-6].

Interface and Communications

In order to learn a set of concepts from an instructor, the agent should be able to communicate with the instructor. This can be done at a lower level: through observation and interpretation of instructor's behavior, or at a higher level. Accordingly, possible agent's interface channels can be divided into lower level channels (e.g., vision, audition, motion control) and higher level channels, for which we assume that semantics of a message can be injected

directly into the architecture. We explained previously [3, 4] that lower-level interface skills are not necessary for the cognitive growth to start: they can be added later or gradually acquired in the process of learning. Therefore, achieving a human-level performance in lower-level interface skills (including visual and auditory input processing, motor output control, speech and gesture recognition, natural language processing) should not be the bottleneck in cognitive architecture design and evaluation. On the contrary, having a human-competitive cognitive core in the embryo architecture [6] is vital for the bootstrapped cognitive growth to start. In particular, we showed in [3, 4] how the paradigms of psychological testing can be adapted for intelligent agents that may lack lower cognitive and behavioral skills, as well as human-level communication abilities. Nevertheless, the ability of an instructor to explain a new concept to the agent is critical and should be supported by the cognitive embryo. Natural language provides a natural, efficient and universal means of communication with a human instructor. Therefore, it is highly desirable, if not necessary, for an efficient universal learner to be able to acquire natural language, starting from limited innate capabilities. The GMU BICA architecture allows for an efficient internal representation of virtually any semantics that can be expressed using natural language [6, 7]. We see a solution in an agent that learns concepts and simultaneously associates with them elements of language in one and the same interactive teaching paradigm.

Cognitive Task and Domain

Implementation of cognitive architectures is the only visible road to computational consciousness. In order to create an implementation of a cognitive agent, one needs to select a life paradigm for the agent: its world, embodiment, needs and desires, expected functional role, etc. When the paradigm is selected, it becomes a task on its own, most efficient solutions for which may not look like anything that we would call alive, genuinely conscious and intelligent, and their implementation may prove nothing new related to computational consciousness. It looks like we have a catch-22 problem. In order to break this cursed circle, we need to reject the paradigm of finding most efficient solutions for well-defined problems as a paradigm of research. Still, we need to select a specific task in order to make a simplest step forward through implementation, but solving the selected task per se should not be the goal (therefore, traditional problem solving criteria for success may not be useful here). The actual scientific goal is *to create and to study the emergent phenomenon of computational consciousness*, regardless of its immediate practical usability. This ideology underlies the selection of examples described below.

GMU BICA: Overview and Future Directions

In order to build a universal learner, our team at George Mason University supported by a DARPA BICA Grant implemented a rapid prototype of the GMU BICA architecture [2-6]. Figure 2 shows a rapid prototype of GMU BICA in action. In this scenario an agent represented by a red dot navigates an indoor environment. It learns features of the environment, their associations with location, as well as possible moves between locations in the environment. From this information the agent builds a cognitive map, so that later on the agent will be able to find a shortest path to a specified feature in the environment. When the acquisition of the cognitive map is completed,

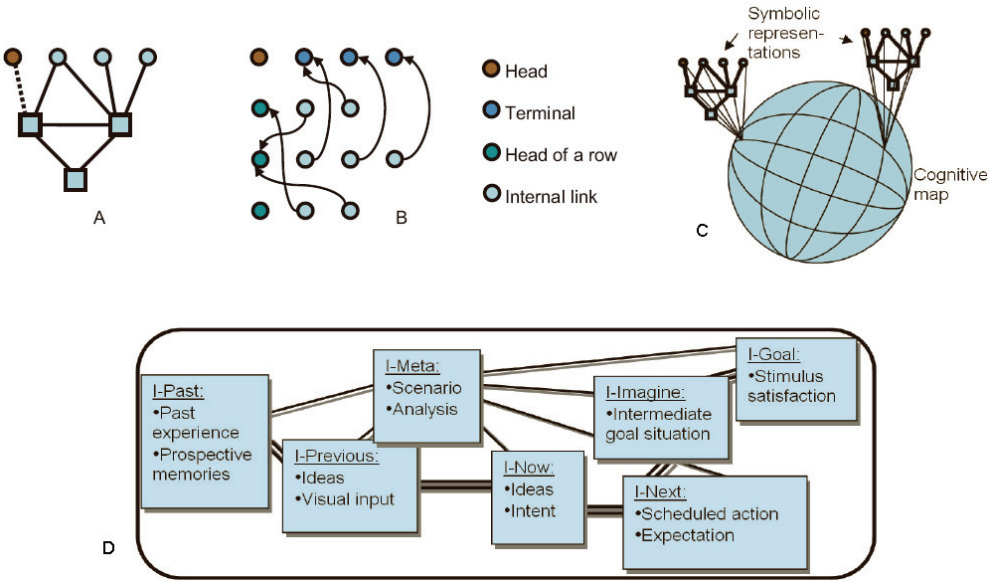

Figure 1. Building blocks of the architecture. A, B: two representations of a schema; C: cognitive map; D: working memory populated by mental states. Episodic memory is organized similarly to working memory, and semantic memory is a collection of schemas.

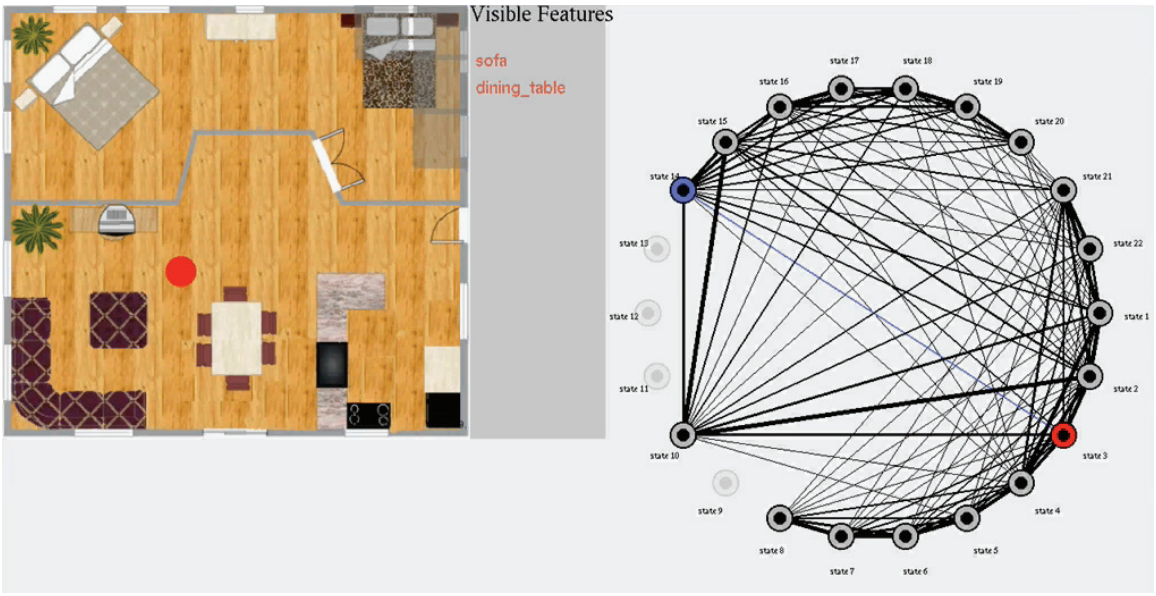

Figure 2. Left: GMU BICA agent (red dot) finds a specified object (sofa) after learning the environment. The task is solved with the help of a cognitive map (right).

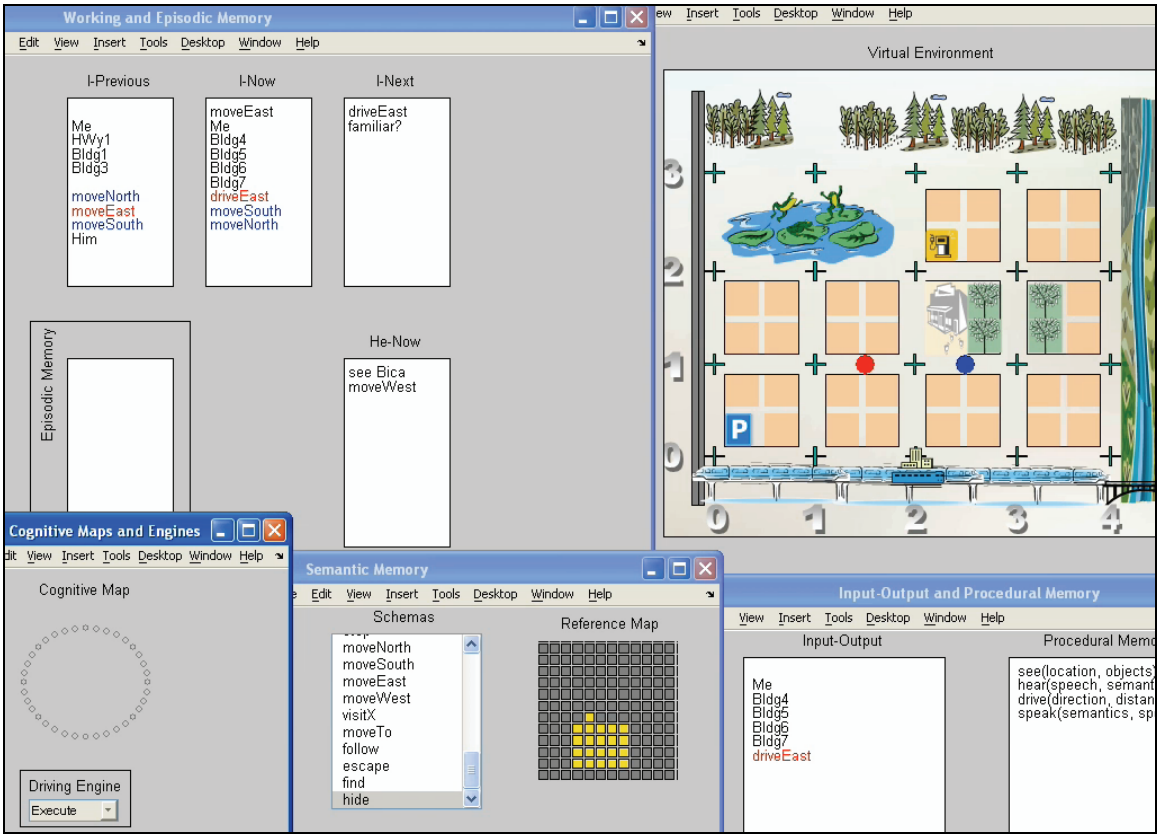

Figure 3. Components of GMU BICA in action: working and episodic memory (top left), the driving engine (bottom left), the environment (top right), the input-output buffer, procedural memory (bottom right) and semantic memory (bottom center).

the agent receives a request to find the sofa. Starting from the bedroom, the agent finds a shortest path to the goal: a location from which the sofa can be seen. All reasoning involved in this task is done with the cognitive map that learns spatial relations among features of the environment.

Figure 3 illustrates dynamics of the architecture components in another paradigm. In the top left corner are two components: working and episodic memory of the agent populated by mental states. Each mental state is shown as a white box populated by instances of schemas: concepts of which the agent is aware in this mental state. Mental states are shifted according to their perspective modification due to the subjective time flow, and eventually they reach episodic memory. The input-output component linked to the procedural memory is shown in the bottom right corner. Instances of schemas populate the input/output buffer. Simultaneously, the related schemas and the corresponding elements of the cognitive map become activated. This is shown at the bottom. In the bottom left corner, alternating states of the Driving Engine that runs all components of the architecture are represented. The virtual environment with the agent in it (the red dot) is shown in the top right corner. Behavioral paradigms gradually change from simple to more complex.

At some point of the scenario the agent discovers a 'playmate': a blue dot that represents another agent in the same environment. Therefore, our BICA agent (the red dot) starts simulating mental states of the other agent and engages in a pretend play, during which it intermittently tries to capture or to escape from the playmate. In this game, the agent learns to recognize intentions from behavior and to achieve goals in paradigms involving other minds.

Figure 4. A bootstrapped learning paradigm. The virtual robot receives verbal instructions and based on them develops new schemas.

Figure 4 is a snapshot of a scenario that demonstrates the advantage of bootstrapped cognitive growth. The idea is to learn new schemas step by step, when more complex schemas are built from previously learned simple ones. The agent learns from human instructions in one simple paradigm, in which the task is to move the box from Roof A to Roof C. As our simulation shows, it is practically impossible to achieve this goal by random exploration of all available actions, no matter how much time is spent. With bootstrapped learning, however, the goal is reached in a few steps. In this imaginary scenario, the robot is instructed by a human and learns action schemas by generalizing human instructions. Each previously learned schema can be invoked by a verbal reference to it and later becomes a part of a new, more complex schema. For example, at the beginning, the robot learns how to go to a certain object and how to move a certain object to a certain place, when there are no obstacles. When this step is completed, the robot can learn how to move the box from one roof to another, using the plank as a bridge. Here the trick is to shift the center of mass of the plank by placing the box on one end of it. The robot does not understand the physics of this procedure, but can learn it from human instructions. Having mastered this action schema, the robot can deliver the box anywhere in the given environment.

In summary, the agent can have a universal learning capability, if it can construct arbitrary schemas in its mind using major paradigms of learning: from instruction, from observation of an example of behavior, from guided exploration, and from analysis of episodic memories. In the above example, the agent learns from instructions using a schema template for representing them. A future direction for GMU BICA should be the addition of limited innate language skills together with the ability to acquire linguistic primitives and associate them with new learned concepts in an interactive teaching paradigm.

Comparison with Related Architectures

Related cognitive architectures, e.g., LIDA [8], CLARION [9] and ACT-R [10], offer similar capabilities, yet in our view they are not sufficient to evolve a computational consciousness: in contrast with GMU BICA, they lack the human-like self [1], and their knowledge representation format imposes severe limitations on learning abilities.

Discussion: An Old Philosophical Problem and Its Proposed Solution

Suppose we have implemented a universal learner. When should it be considered conscious? It seems that the boundary between Nature and artifacts, between neuronal and silicon bases of information processing is fuzzy. For example, when a speaker at a neuroscience conference describes a modern experimental setup, one may not notice the moment when (s)he stops talking about information flow in a biological substrate and starts describing information flow in a silicon or metallic substrate, or vice versa. Where is the real boundary of the substrate of cognition? What parts of the system may have subjective experiences? Let us look at a more general problem. Our attribution of subjective experiences to other people is arbitrary and subjective. Their actual experience may be different or may not be present, or may be present in 50% of all cases, etc., with no consequences for any objective measurement conducted within the framework of modern empirical science (as pointed by Chalmers [11]), because this science does not have a room for subjective experience per se as a subject of study. The questions "how does an object look from an outside" and "how does it feel from

inside" are not reducible to each other. E.g., for me the fact that I happened to have a "feeling from an inside" is not an illusion. But I cannot infer from any present scientific fact that the same statement about another brain is not an illusion. There are at least two possibilities left, and nobody can resolve them by measurements within the modern scientific paradigm.

It appears, however, that the Hard Problem [11] understood in this sense can be solved in exactly the same way as any of the "Easy Problems": using the classic paradigm of empirical science, although on a new turn of the spiral of evolution, with the new science of the mind that has to be created. The job is not done automatically by describing neuronal dynamics of the brain in every detail, from the molecular level to the system level. Moreover, it is not sufficient to stipulate that consciousness emerges at a certain level of complexity [12]. To solve the problem, it is necessary to treat experiences as an element of Nature and a (new) subject of study, to introduce new axioms, new paradigms of measurement and new metrics for them.

The "old" scientific method, no matter how rigorous and well-grounded it seems, is in fact based on a set of beliefs: (i) that experiences reflect elements of reality; (ii) that reality obeys simple laws, and (iii) that one can develop general knowledge by empirically testing parsimonious hypotheses. The new turn of the spiral follows the same logic, starting with an observation that there is a different kind of experience: "conscious higher order thoughts" [13, 14] that convey information about experiences themselves. Accepting this fact of observation, we can postulate that (i) higher order thoughts reflect our experiences that are elements of reality; (ii) experiences obey universal laws that can be learned and stated as theories, (iii) by the parsimony principle, physical systems with similar functional organization must have similar experiences. The last axiom is known as the principle of organizational invariance (POI [11]). It allows us to make a connection between the two sciences – and between the two kinds of consciousness, too. No further "proof", "explanation" or reduction of consciousness is necessary.

From this point of view, a universal learner becomes a computational consciousness when it reaches a level of functional organization that is a characteristic of the human mind. In human development, there is a connection between first memories of conscious experiences and development of the self. Therefore, the presence of a human-like self [1] cold be a criterion for consciousness that is applicable to artifacts. A functionalist model of the human-like self is built into GMU BICA; therefore, a universal learner based on this architecture can be expected to develop full functionality of the self by learning (about its own mind rather than about its body or environment).

A necessary fundamental tool suggested by this logic is a metric system to be used for semantics of subjective experiences, and here our semantic cognitive map concept comes to the rescue. The underlying general idea is that it is possible to represent semantic relationships among symbolic representations quantitatively, using geometrical concepts such as distances and angles. Generally, a semantic cognitive map of a given representation system can be defined as an abstract metric space onto which the given set of representations is projected, such that the dimensions and the metrics of the space reflect semantic relationships among the representations (compare with the qualia space concept discussed by Tononi and Edelman [12]). These representations can be, for example, schemas or words of natural language. The idea is not new: it was considered in cognitive psychological literature and dismissed a long time ago [15]. Nevertheless, the notions of a semantic space and semantic cognitive mapping remain a hot topic [7, 16] and acquire rapidly increasing popularity.

In summary, here is the answer to the question at the top of this section. The axiom of POI together with the self concept defined within the new science of mind will allow us to detect a point at which a universal learner will develop a self with its own, computational consciousness.

Conclusions

Imagine that you sit at your own computer, open a window of an application, and type: "I want you to write a program that computes the number of topologically distinct trees of *N* nodes". What will happen? Either nothing or you will get an error. It would be a miracle if, as a result, you would see a Java code that solves the problem. This miracle can happen, if one starts with the right agent architecture and teaches it step by step, and the first task that this agent would solve could be much more primitive than the above example.

In summary, the field of computational consciousness can be contrasted with the field of problem solving in AI along several lines.

- A goal in problem solving is to solve problems. Success is measured in the number and the quality of solutions. In computational consciousness, the selected problem and the quality of its solution are relevant to the goal of computational consciousness no more than the selected piece of paper and the quality of the ink are relevant to the content of the letter. The goal in computational consciousness research is to create and to study the phenomenon of computational consciousness.
- A goal in problem solving is to integrate new developed tools and to grow powers of machine intelligence bigger and higher. In contrast, a goal in computational consciousness is to come up with each new cognitive embryo design smaller and more primitive than its predecessor, separating the vital cognitive core from peripheral components that are not critical, while maintaining the requirement that one embryo should potentially be able to solve all problems in the world.
- Intelligent agent architectures intended for problem solving are designed to serve specific purposes in specific paradigms. In contrast, the goal and the paradigm of a particular human life are uncoupled from

the human evolutionary origin and genetic design. For example, a function of a particular human life could be to create a quantum computer, while this goal would not make any sense in the epoch when the human genome evolved. The fact that a human can, in principle, independently choose and achieve higher goals that are not a part of the human design separates human consciousness from lower animals and from automata. Therefore, if we are ever to create computational consciousness, we must be able to replicate this same feature in a machine. A way to do this is by starting with a universal learner designed as a parsimonious cognitive embryo representing a critical mass of learning capabilities.

Acknowledgments

I am grateful to the members of our team that together with me designed and implemented rapid prototypes of GMU BICA: Dr. Kenneth A .DeJong, Dr. Giorgio A. Ascoli, Mr. Mark A. Coletti, Mr. Robert Lakatos and Mr. Deepankar Sharma, being supported by a DARPA BICA Grant "Integrated Self-Aware Cognitive Architecture". I thank Dr. Kenneth A. DeJong for suggestions how to improve the presentation, which were very useful here and will be fully addressed elsewhere.

References

[1] Samsonovich, A. V., and Nadel, L., 2005. Fundamental principles and mechanisms of the conscious self. *Cortex*, 41 (5): 669-689.

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

[3] Samsonovich, A. V., Ascoli, G. A., and De Jong, K. A., 2006. Computational assessment of the 'magic' of human cognition. In *Proceedings of the 2006 International Joint Conference on Neural Networks*, pp. 1170–1177. Vancouver, BC: IEEE Press.

[4] Samsonovich, A. V., Ascoli, G. A., and De Jong, K. A., 2006. Human-level psychometrics for cognitive architectures. *Fifth International Conference on Development and Learning ICDL 2006, Bloomington, IN, Department of Psychological and Brain Sciences, Indiana University.* CD-ROM, ISBN 0-9786456-0-X.

[5] Samsonovich, A. V., Ascoli, G. A., De Jong, K. A., and Coletti, M. A., 2006. Integrated hybrid cognitive architecture for a virtual roboscout. *Cognitive Robotics: Papers from the AAAI Workshop.* M. Beetz, K. Rajan, M. Thielscher and R. B. Rusu. Menlo Park, CA, AAAI Press. WS-06-03: 129-134.

[6] Samsonovich, A. V., 2006. Biologically inspired cognitive architecture for socially competent agents. *Cognitive Modeling and Agent-Based Social Simulation: Papers from the AAAI Workshop.* M. A. Upal and R. Sun. Menlo Park, CA, AAAI Press. WS-06-02: 36-48.

[7] Samsonovich, A. V., and Ascoli, G. A., 2007. Cognitive map dimensions of the human value system extracted from natural language. In B. Goertzel and P. Wang (Eds.). *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms. Proceedings of the AGI Workshop 2006. Frontiers in Artificial Intelligence and Applications,* vol. 157, pp. 111-124. IOS Press: Amsterdam, The Netherlands.

[8] Franklin, S., 2007. A foundational architecture for artificial general intelligence. In B. Goertzel and P. Wang (Eds.). *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms. Proceedings of the AGI Workshop 2006. Frontiers in Artificial Intelligence and Applications,* vol. 157, pp. 36-54. IOS Press: Amsterdam, The Netherlands.

[9] 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.

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

[11] Chalmers, D. J., 1996. *The Conscious Mind: In Search for a Fundamental Theory.* New York: Oxford University Press.

[12] Tononi, G., and Edelman, G. M., 2000. *A Universe of Consciousness: How Matter Becomes Imagination.* New York: Basic Books.

[13] Rosenthal, D. R., 1993. Multiple drafts and higher-order thoughts. *Philosophy and Phenomenological Research*, LIII: 4, 911-918.

[14] Rosenthal, D. R., 2006. Consciousness and intrinsic higher-order content. In Hameroff, S., et al. (Eds.). *Tucson VII: Toward a Science of Consciousness. Consciousness Research Abstracts: A Service From Journal of Consciousness Studies*, program no. 39, page 52. Tucson, AZ: Imprint Academic (online supplement at http://davidrosenthal1.googlepages.com/Tucson-7-Powerpoint.pdf).

[15] Tversky, A., and Gati, I., 1982. Similarity, separability, and the triangle inequality. *Psychological Review* 89 (2): 123-154.

[16] Gärdenfors, P., 2004. *Conceptual Spaces: The Geometry of Thought.* Cambridge, MA: MIT Press.