

Meta-Cognitive Architecture for Team Agents

Alexei V. Samsonovich (asamsono@gmu.edu)

Krasnow Institute for Advanced Study, George Mason University
4400 University Drive MS 2A1, Fairfax, VA 22030-4444 USA

Kenneth A. De Jong (kdejong@gmu.edu)

Department of Computer Science and Krasnow Institute for Advanced Study
George Mason University, 4400 University Drive, Fairfax, VA 22030-4444 USA

Abstract

A key element of our approach is the interpretation of “self” in a meta-cognitive sense: that is, “self” is understood as a virtual character representing an agent as the subject of experience, as the target of attribution of experiences and deliberate actions performed by this agent. Thus understood, “self” can be represented as an element in an agent’s cognitive system and can be used for meta-cognitive processing: i.e., reasoning about one’s own self and other selves. This general idea reflects a simulationist theory-of-mind viewpoint (Nichols & Stich, 2000), which is taken as the basis for our approach. Our model of an agent’s mind includes multiple instances of “self” representing notions of I-Now, I-Yesterday, I-Imagine, I-Goal, etc. Each instance of “self” is represented by a “chart” with a set of properties and mental states attributed to it. Thus, mental states in this framework are representations of experiences attributed to a particular instance of “self”. This attribution further implies certain rules and constraints imposed on the contents and the dynamics of representations. The result is a general architecture that will enable in intelligent agents a meta-cognitive “common sense”, which proves to be vital in a variety of paradigms and scenarios requiring cooperation within a team.

Introduction

The area of cognitive teams requires a new approach for creating and managing multi-agent systems that can cooperate, make joint decisions and achieve goals in a coordinated way. Such systems can include robots, virtual agents and people. The task of achieving successful performance of an ad hoc team or network of intelligent agents requires a new technology that would enable self-awareness, meta-cognition, introspection, theory of mind, episodic memory and “what if” capabilities in agents. All these abilities are closely related to the notion of a “self”.

Building artificial cognitive systems that possess a concept of “self” and a notion of “self-awareness” is a difficult task, but one with tremendous practical significance and potential with the increasing interest and emphasis on autonomous, agent-based systems. As we lay the groundwork for teams of robots, agents and people, it is difficult to imagine effective team dynamics without individual team members having a sufficient sense of “self” that allows for self-reflection, self-assessment, self-improvement, and understanding of others. In the present

work we outline a means for achieving this goal by bringing together recent developments in cognitive science and in artificial intelligence: specifically, by relating to each other the results of theory of mind studies (Goldman, 2000; Nichols & Stich, 2000), of cognitive modeling based on production systems (Soar: Newell, 1990; ACT-R: Anderson, 1993; Anderson & Lebiere, 1998), intelligent agent architectures (belief-desire-intention, or BDI: Bratman, 1987; Ortiz, 1999; Dix et al., 2001; Sabater et al., 2002; Panzarasa et al., 2002; Dragoni et al., 2002), and team robotics in situated contexts (Tambe, 1997) of team problem solving such as search-and-rescue scenarios.

The design of a meta-cognitive architecture outlined in this work allows for its implementation as software that could be installed in virtual agents, as well as in mobile robot agents, in a form that can be used by individual agents for a variety of cognitive tasks requiring cooperation within a team. We argue by several example scenarios that, in order to be successful in real life situations, the agents in the team must possess meta-cognitive awareness of self and others, episodic memories, and the ability to explore plausible “what if” scenarios by mental simulations.

Specifically, the following abilities will become available in agents based on the proposed architecture: (a) the ability to be aware of self at a meta-cognitive level, meaning awareness of current self actions and mental states, goals and intentions, of ones personal role in the global scenario, and the ability to evaluate own behavior; (b) the ability to understand a global picture involving multiple agents and their mental states via a simulationist theory-of-mind approach; (c) the ability to remember previous experiences, to be aware of them and to find among them relevant aspects to apply to the current situation (episodic memory, learning from personal experience); (d) the ability to explore via mental simulations “what if” scenarios and to plan actions based on these simulations.

In addition, the proposed cognitive architecture is sufficiently general and powerful to support the following features: (e) the ability to explain own behavior and to accept directions, using human-level communications; (f) the ability to mentally simulate emotional states in order to better understand and serve human agents; (g) the ability to learn new general concepts from experience and/or interactions with a teacher; and (h) the ability to exhibit “voluntary” meta-cognitive rational initiative in order to

improve performance via re-designing itself. The design and implementation of these additional features will become possible when, in the near future, solutions to the related lower-level problems are available, e.g., the problem of natural language processing.

Technical Approach

The key element of our approach enabling the new meta-cognitive abilities in agents is a framework in which the concept of "self" and associated mental states are viewed and represented as outlined below (see also Samsonovich & Nadel, 2003; Samsonovich & De Jong, 2003).

Conceptual Framework

In this framework each instance of the self is characterized by its unique mental perspective, including the identity of the agent (e.g., "I"), the time stamp of the associated experience (the "subjective time"), the status of the instance (e.g., actual, imaginary, remembered), the position in the theory-of-mind hierarchy (Baron-Cohen, 1995; Nichols & Stich, 2000), and a general context in which the experience occurs (e.g., a spatial location). Instances of "self" are labeled in this text accordingly: I-Now, I-Previous, I-Next, I-Past, I-Imagined, etc. Thus, instances of the self are introduced as abstract entities possessing a set of parameters and properties attributed to them, but lacking any internal structure or mechanisms. In general, a transition from an ordinary cognitive representation framework to a system of mental states involves (i) partitioning the working memory by a set of instances of the self and (ii) imposing certain constraints and rules based on this partition. Each instance of the self represents one mental perspective and is associated with a domain of the partition called here a "chart". The dynamics of mental states on one chart is called here a mental simulation. In our framework, multiple charts together with their associated mental simulations (each with its own instance of "self") may be co-active in the same cognitive system at any moment of time. Together they constitute working memory of the system.

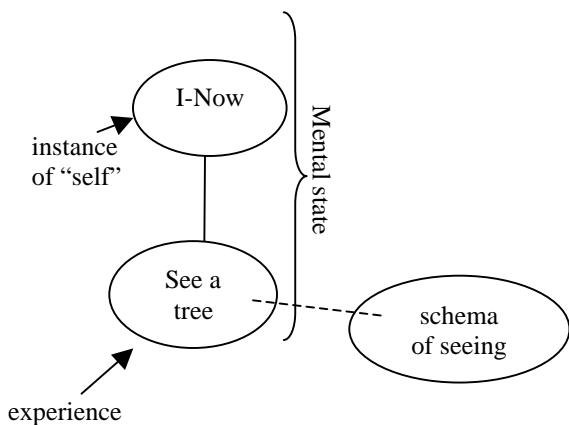


Figure 1: An example of a mental state.

An example of a mental state could be a representation of an action attributed to I-Now that is formally interpreted by the system as an action caused by the agent at the current moment of time. Another example could be a representation of an image attributed to I Previous, that is formally interpreted by the system as an image subjectively perceived by the agent at a previous moment of time. Generally, a mental state is a pair of a cognitive representation (understood in a broad sense) and an instance of the agent's "self".

All cognitive representations in this framework are created based on schemas. The term "schema" (plural "schemata" or "schemas") was introduced by Kant (1781/1929), and currently has an extremely broad usage in science, with different semantics in different fields. Within computer science alone, the word has perhaps a dozen different senses. An advanced theory of schemas can be found in evolutionary computation (Langdon & Poli, 2002). In this text, the notion of a schema is understood in a very general sense applicable to cognition. We define a schema as an abstract model or a template that can be used to instantiate and to process a certain mental category. The categories may include all possible types of elements of the subjective world: concepts and beliefs (e.g., objects, properties, events, relations), feelings, sensations, intentions, actions, etc.

We generally say that a schema has a state, when its instance is bound to some given content. For example, seeing a red circle can be described via a state of the schema of red. Logical reasoning can be described in terms of states of the schemas of inference, etc. In our terminology, a state is considered "mental", if it is attributed to a subject of experience, i.e., to a "self". In our framework, schemas are dynamical objects: they can be created and modified "online", and they all are represented in one universal format. The entire set of schemas in a given individual constitute that individual's semantic memory, i.e., the general knowledge about self and the world.

In addition to being attributed to a particular mental perspective (e.g., I-Now), each mental state is characterized by an attitude. The word "attitude" here refers to a kind of a functional role of a state on a chart with respect to other content and the instance of the subject's self. Examples of attitudes are: intended, desired, believed, dreamed, recalled, "my own", somebody's, etc. In other words, the attitude characterizes the kind of a mental position of the subject with respect to the content of the mental state.

Architecture

The macro architecture of an individual agent includes the following components: the input and output buffers, working memory, episodic memory, goal-and-plan memory, semantic memory, and procedural memory. The input-output buffers are special charts labeled I-Input and I-Control, that can be formally considered as a part of the working memory. Episodic memory consists of selected previously active charts that became de-activated and stored in a long-term memory together with all their mental states.

Goal-and-plan memory is a counterpart of the episodic memory and is similar to it, except that it consists of selected previously active simulations of imaginary scenarios and goal situations, together with a certain system of values. The semantic memory consists of a set of schemas, and the procedural memory consists of a set of drivers: these elements are defined below.

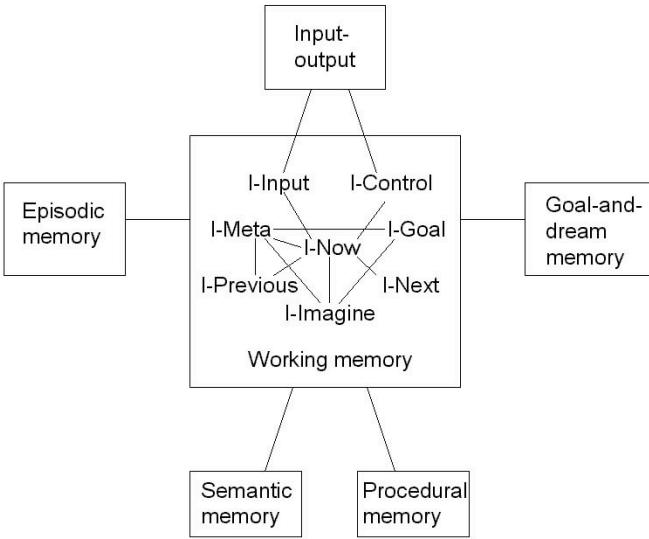


Figure 2: Macro-architecture of the system.

Computational Format

From a computational point of view, we can say that our notion of a schema generalizes the notion of a production in Soar and the notion of a chunk in ACT-R: schemas and mental states are data structures rather than active elements. They need drivers in order to function.

Here by a “driver” we refer to an active object (e.g., an executable function) that performs standard procedures of processing of charts representing mental perspectives, schemas, mental states, and the relations among them. Procedures are separate elements of our framework: they are scripts associated with drivers that represent basic metacognitive skills. Drivers and procedures constitute the content of the procedural memory.

In particular, procedural memory in this framework includes the following drivers: Clock (chart status updating and subjective time flow), Predictor (probing candidates for mental states), Scanner (binding schemas), Completer (executing states), Terminator (eliminating states to keep pre-set memory limits for each chart), Stimulator (goal activation), and Ego, that performs a broad spectrum of tasks at a meta-cognitive level (voluntary actions, mental simulations, internal conflict resolution, self-evaluation, etc.). All drivers may work in parallel. In addition, most drivers may exist in multiple copies working in parallel. This circumstance makes the model suitable for implementation on parallel computers.

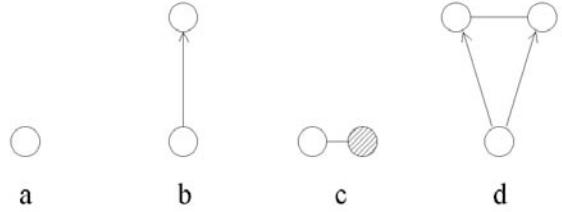


Figure 3: Simplest geometrical configurations of schemas: an entity (a), a property (b), an event (c), and a relation (d).

Again, states in our framework are bound instances of schemas. One way to represent a schema is to view it as a graph, the nodes of which are associated with “terms”. Each term represents some mental category (and therefore refers to a schema associated with that category). The root term of a schema represents the mental category associated with this schema itself. A simplest design of a schema is just the root term. Each term, either in a schema or in a state, is an object (in the object-oriented-programming sense) with a standard set of slots that specify parameters of the term, including the name of a mental category, a mental perspective, an attitude, the mode and the status of binding, etc. To bind a schema means to assign particular values to parameters of its terms (not necessarily all parameters and all terms). Generally, a schema specifies constraints and relations among the values of parameters of its terms. In addition, it may specify the order in which the terms should be bound and “side effects” of binding: e.g., creation of related states, schemas, etc.

A chart can be viewed as a container in which a particular mental simulation takes place. It also provides a label (e.g., “I-Now”) attached to all elements of this mental simulation, and a domain of the system’s cognitive space with individual locations in it understood as mental attitudes of the instance of “self” associated with the chart. In addition, each chart is characterized by its relations to other charts and its position in the global theory-of-mind hierarchy. These relations determine the rules of information exchange among charts, which is implemented based on messages. A message is another key element of this framework. Each message is characterized by two mental states: the source and the target. One nice feature of this framework is that the same format that is used for internal messaging can be used for communications among agents in a team.

Example Scenario of a System in Action

Any intelligent collaboration within a team of intelligent units requires understanding other minds. Regardless of whether a task is to carry a heavy object by joined efforts, to keep each other informed about critical, locally available information, or to perform a joined maneuver in capturing an enemy, it is not possible to be successful in own personal role without relying on the predictability of partner’s behavior. Therefore, members of the team must possess awareness of the partners’ internal states and cognitive abilities. Similarly, they must possess a concept of self in order to understand their own personal role, their own mental state and their own cognitive abilities. One possible

alternative to this scheme could be to turn local intelligence off and to rely on a centralized control. However, in most scenarios typical for military operations, search and rescue operations and the like, it is not always safe or even possible to use centralized control of a mission performed by an ad hoc, heterogeneous team of robots, software agents and people in a hostile environment. Units of the team (including those performing centralized control) could be lost or damaged, and global communications could be discontinued for a variety of reasons. In addition, there must be technological and common-sense limitations on the frequency of long-range communications: e.g., agents probably should not continuously broadcast all their video and other sensory input. Therefore, intelligence in the team must be distributed, with certain rules of subordination, ethics, etc. Individual team members must have a means of local decision making based on an understanding of the minds of their partners and themselves. In addition, they must be able to learn from personal experience, to reason about possible future scenarios involving the team, and to quickly and robustly respond to surprises. In the following example scenario we demonstrate why these features may become vital, and why their “traditional” implementation, e.g., based on mathematical logic, may not be acceptable.

Consider a possible scenario in which two agents that perform a collaborative surveillance task capture an intruder. They get him to cooperate using verbal commands and warning gun shots. Suppose that, after a warning shot given by chance simultaneously by both agents, the intruder falls down and shows no signs of life. An immediate account of this event given by each agent could be that their partner killed the intruder. Few seconds later, another intruder opens fire on the agents from the opposite side. The agents have to turn around and to respond with their guns. At this moment the first intruder throws a hand grenade that destroys them both. This would not happen if one of the agents kept an eye on the first intruder considering the possibility that he was still alive. Robots capable of logical reasoning may not succeed in this story, given that they have virtually no time for extensive reasoning or for communications. In order to succeed, the agents need to understand each other and to be able to think in terms of a pretend-play schema attributed to the intruder.

And the Miracle Happens...

In order to be more specific, and to demonstrate the advantages of the proposed meta-cognitive architecture in a trivial virtual reality paradigm, we present a simple computer simulation result. The paradigm is that two agents, Circle and Triangle, are placed in a rectangular area, which has a bridge attached to it (Figure 4 a). The bridge is implemented as a twisted belt, a part of a Möbius strip, which makes the entire manifold non-orientable. There are seven letters randomly allocated in the area: M, I, _, A, C, L, and E. The agents can see and recognize letters, including their orientation (one letter, _, is mirror-reflected). Agents cannot flip letters. The task for the team is to spell the word “MIRACLE”, with all letters upright, relying on a mental

representation (template) of the goal that they both have stored in their semantic memory.

There are only two possible elementary physical actions that an agent can perform in this world. (1) Pick a letter and put it in its proper position in the virtual template, if this is possible, and if this is not possible, hold it. (2) Pass over the bridge. Other possible (cognitive rather than physical) actions include the following. (3) Mentally allocate a goal template in the environment. (4) Send a message to the partner (see below). In addition, the agents can perform a number of meta-cognitive actions, as described below. Plus, they have a concept of a horizontal flip of any object (including self), but cannot perform flips. All details of spatial navigation and letter manipulation are presumed to be implemented at a lower (automated) level and do not enter the agents’ minds (e.g., agents do not have a concept of spatial coordinates). The time is discrete, and an agent can perform at most one physical action at a moment of time.

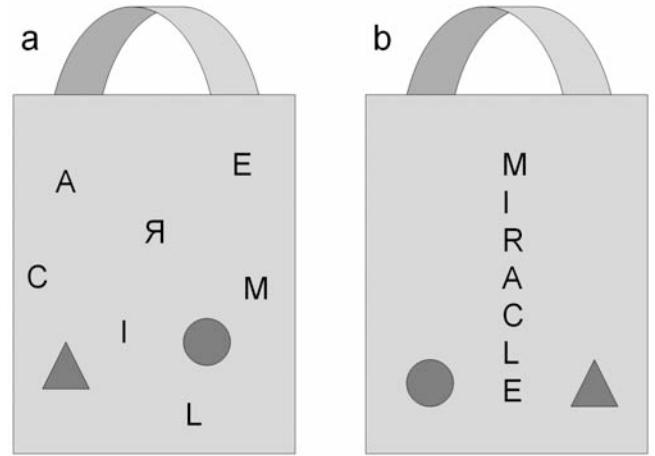


Figure 4: A virtual reality simulation example.

Agents communicate with each other by transmitting any number of selected own mental states to the partner. This is done mostly automatically, when an agent intends to act or becomes aware of something new. In addition, all new schemas (e.g., hypotheses) are shared instantly. Transmitted mental states get represented in the partner’s mind as mental states attributed to the sender. Thus, communications amount to direct copying of internal representations of mental states: this simple choice implies absolute trust and absolute sincerity in the “relationships” among the agents.

Agents do not possess advanced built-in reasoning abilities and follow simple rules. At each step under normal conditions, an agent imagines possible actions (i.e., generates ideas), simulates expected results of imaginary actions, and checks whether the idea of the first move is good in the sense that it results in implementing goal elements. A good idea is accepted as intent, which is communicated to the partner and then executed, if there are no conflicts. When there are no good ideas, agents try any possible actions provided they do not destroy implemented goal elements. This strategy is selected here for its

simplicity and may not work in a more general case; still, the main interest in this simulation study was not in cognitive, but in meta-cognitive dynamics of the system, which we describe next discussing a conflict situation as an example.

A conflict situation emerges when internal mental representations of an agent become inconsistent with each other: e.g., the expected and the actual experience do not match. In this case an agent uses its episodic memory, trying to account for the mismatch. It attempts imagining all possible unknown events that could happen in the environment recently (in our oversimplified world meaning flips of anything, when the agent was passing the bridge). Next, given a parsimonious account of this sort, the agent starts making hypotheses as to the possible causes of the unknown events, trying to interpret the latter as side effects of known physical events. During this process, agents may simulate own and each others' minds, following the principle of parsimony: simplest hypotheses are explored first. When a hypothesis about a side effect is found that resolves the conflict, it is accepted as a new semantic knowledge (schema) that yet has to be tested behaviorally. Based on the new knowledge, agents may change the status of their present and past mental states (e.g., I-Now may become I-False-Belief, and I-Meta, where the analysis was performed from a "third-person perspective", may take the position of I-Now, etc.). In addition, the agents in a case like this are likely to revise their intents, scenarios and, possibly, goals. Finally, the new knowledge is used in the normal process of planning.

This model was implemented and simulated in a computer, producing the result shown in Figure 4 b. The agents were acting in turn. First, Circle allocated a virtual template for the goal, which was accepted and shared by Triangle. Then agents started filling the template with available letters. Because the letter 'R' was not available, Circle decided to cross the bridge. This immediately resulted in a conflict situation, successfully resolved by the system, as it follows from the following output that represents the content of two consecutive messages sent by Circle to Triangle:

Message 1: I am surprised to see R upright, and C, L and E flipped.

Message 2: As a parsimonious explanation, I suspect my own flip that happened when I was passing over the bridge. I hypothesize that the bridge causes flipping.

After this discovery, the hypothesis together with the pre-existing concept of a flip was used in planning by both agents, and the goal was quickly achieved (Figure 4 b).

This simulation result clearly demonstrates that the agent was able to solve a nontrivial puzzle with a single meta-cognitive act of imagining its own flip that could take place in the past. After detecting a conflict between its internal representations, Circle started imagining all possibilities, including various flip events that could happen recently,

until it imagined its own flip. In order to do this, the agent took a third-person perspective I Meta and considered its current instance of self "from an outside". It could take a substantial amount of traditional logical reasoning to do the same (e.g., questioning the semantics of many agent's mental representations would be necessary, and the analysis would be difficult to conduct within the same mental perspective). On the other hand, an elementary meta-cognitive act allows the agent to see immediately "what is it like to be flipped" and to compare this vision to the current experience.

The particular observed outcome was not "pre-programmed" and was not the only possible course of action. For example, given the same initial scenario (including the first message), it would be likely for Triangle to guess independently that Circle got flipped, and to come up with an equivalent hypothesis first.

Connections to Modern 'Hot Topics'

Cognitive Psychology and Philosophy of Mind

Currently there are two main competing points of view on human theory of mind: "theory-theorist" and "simulationist" (Goldman, 2000). The former assumes that people represent mental states in themselves and in others by making inferences from common-sense concepts, while the latter assumes that people use their first-hand experience to understand other minds, in other words, perform "mental simulation".

Our framework viewed as a model of the human theory of mind falls into the simulationist camp. Perhaps, the closest to it is the framework proposed by Nichols and Stich (2000). Specifically, their notion of a Possible World Box (PWB) is similar to our notion of a chart, although PWB does not represent an instance of a self, and therefore does not explicitly provide a means of separating mental states based on their attribution to different instances of the subject's self. Other elements of the model of Nichols and Stich (2000) also can be mapped onto our framework: e.g., their UpDater can be related to a subset of our drivers.

Remarkably, Nichols and Stich emphasize the advantage of the anthropomorphic approach to self-monitoring over logical reasoning: *"When normal adults believe that p, they can quickly and accurately form the belief I believe that p; when normal adults desire that p, they can quickly and accurately form the belief I desire that p; and so on for the rest of the propositional attitudes. In order to implement this ability, no sophisticated Theory of Mind is required"* (Nichols & Stich, 2003).

In addition, the proposed framework can be used in the field of cognitive psychology of neurological disorders to give an account to major agency disorders, including various forms of hippocampal amnesia, various aspects of schizophrenia, multiple personality, PTSD, and autism (Samsonovich & Nadel, 2003). This topic, however, is beyond the scope of the present paper.

Artificial Intelligence

During recent years a tremendous progress has been made in several fields related to intelligent agents possessing meta-cognitive abilities (e.g., Panzarasa et al., 2002). These fields primarily consist of logical foundations of artificial intelligence and of practical approaches based on production (or rule-based) systems. The state-of-the-art intelligent agent architecture based on the BDI framework (Bratman, 1987) and its variations allows for implementing a theory of mind in an agent using logical reasoning. An implementation of this sort would fall into the "theory-theorist" division (see above). It would not, however, provide a natural way of solving unexpected real-life situations like the examples considered above.

In addition, unlike its well-known analogs, e.g., Soar (e.g., Laird et al., 1987) and ACT-R (Anderson & Lebiere, 1998), the system that we propose to build possesses universality and an unlimited potential of development and generalization on the part of its meta-cognitive abilities.

Discussion of Further Perspectives

We expect the following impact of the proposed architecture on the team agent technology and beyond. (1) The software agent architecture will be used in teams of mobile robots. (2) The new technology of meta-cognitive systems will become widely practically available. (3) The results will open a broad field of new possibilities in research, allowing for a "quantum leap" from the existing state-of-the-art technologies (Soar, ACT-R and BDI agents). The set of potentially available agent abilities will include the ability to explain own behavior and to accept directions, using human-level communications, the ability to mentally simulate emotional states in order to better understand and serve human agents, the ability to learn new general concepts from experience and/or interactions with a teacher, and the ability to exhibit meta-cognitive rational initiative and to improve performance via re-designing itself. (4) The theoretical model of a meta-cognitive system will be related to human cognition and to the functional organization of the human brain, thus resulting in a qualitatively new cognitive-psychological model. Mapping of model components onto the neuro-anatomical organization of the brain will allow for a model-based interpretation of "mysterious" neurological disorders, as well as for a better understanding of a normal state of the human mind. (5) Finally, a field of computational consciousness will be brought to existence.

Despite many recent speculations, the field of computational consciousness has not been born yet. In our view, the idea behind its likely origin is that today we may be in a position to create a new object of study for such abstract disciplines as philosophy of mind and psychology of higher cognitive functions. Rather than creating a computer simulation of another abstract and oversimplified cognitive-psychological model of human mind, our present ambition is to create a virtual entity emulating human mind in its most essential abilities, an entity that by itself might be of a great scientific and practical interest for us. This entity might further require its own theoretical explanation rather

than provide an immediate account of human cognition, however, that future theory, when found, might eventually help us to understand our own mind.

References

Anderson, J.R. (1993) *Rules of the mind*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J.R., & Lebiere, C. (1998) *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum Associates.

Baron-Cohen, S. (1995) *Mindblindness: An essay on autism and theory of mind*. Cambridge, Massachusetts: Bradford Book, MIT Press.

Bratman, M.E. (1987) *Intentions, Plans, and Practical Reason*. Harvard University Press: Cambridge, MA.

Dix, J., Kraus, S., & Subrahmanian, V.S. (2001) Temporal agent programs. *Artificial Intelligence* 127: 87-135.

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

Goldman, A. (2000) Folk psychology and mental concepts. *Protosociology* 14: 4-25.

Kant, I. (1781/1929) *Critique of pure reason*. Translated by N. K. Smith. New York: St. Martin's Press.

Langdon, W.B., & Poli, R. (2002) *Foundations of genetic programming*. Berlin: Springer.

Laird, J.E., Newell, A., & Rosenbloom, P. (1987) Soar: an architecture for general intelligence. *Artificial Intelligence* 33: 1-64.

Nichols, S., & Stich, S. (2000) A cognitive theory of pretense. *Cognition* 74: 115-147.

Nichols, S., and Stich, S. (2003) Reading one's own mind: A cognitive theory of self-awareness. In Smith, Q., and Jokic, A. (Eds.), *Aspects of Consciousness* (in press), Oxford, UK: Oxford University Press.

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

Sabater, J., Sierra, C., Parsons, S., & Jennings, N.R. (2002) Engineering executable agents using multi-context systems. *Journal of Logic and Computation* 12 (3): 413-442.

Samsonovich, A.V., & Nadel, L. (2003) *Fundamental Principles and Mechanisms of the Conscious Self. Cortex: Special Issue on Brain, Mind and Consciousness* (under review).

Samsonovich, A.V., & De Jong, K.A. (2003) Definition of a computational framework for meta-cognitive systems. *Cortex: Special Issue on Brain, Mind and Consciousness* (under review).

Tambe, M. (1997) Towards flexible teamwork. *Journal of Artificial Intelligence Research* 7: 83-124.