International Journal of Machine Consciousness
Vol. 1, No. 1 (2009) 111–130
© World Scientific Publishing Company

THE MENTAL STATE FORMALISM OF GMU-BICA

ALEXEI V. SAMSONOVICH

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

KENNETH A. DE JONG

Computer Science Department and Krasnow Institute for Advanced Study,
George Mason University, 4400 University Drive MS 2A1,
Fairfax, Virginia 22030-4444, USA
kdejong@gmu.edu

ANASTASIA KITSANTAS

College of Education and Human Development,
George Mason University, 4400 University Drive MS 4B3,
Fairfax, Virginia 22030-4444, USA
akitsant@gmu.edu

GMU-BICA, the biologically-inspired self-aware cognitive architecture developed at George Mason University, continues to be a useful prototype for various intelligent artifacts, including intelligent tutoring systems, yet the underlying formalism of mental states used in its design was never described in detail. The present theoretical work aims at filling this gap, focusing on the top cognitive level (mental states) and leaving detailed description of the lower level (schemas), as well as non-declarative components, for future publications. Among the distinguishing features of the GMU-BICA mental state formalism are: (i) a subject-centered view of the world, (ii) the multiplicity of mental perspectives simultaneously represented in working memory, each playing its unique functional role, and (iii) the limited span of awareness. This model is consistent with human psychology and gives testable predictions. The work explains, through analysis of examples, how the framework can be used to build computational models of self-regulated learning, and why in this case it is expected to unleash a new for artifacts power of human-like cognition and learning.

Keywords: Cognitive architectures; metacognition; self-regulated learning; human-level AI.

1. Introduction

The purpose of this work is two-fold: (a) to explain the formalism of mental states developed as a part of the DARPA IPTO BICA project at George Mason University ("BICA" stands for "Biologically Inspired Cognitive Architectures"), and (b) to connect this formalism with a model of self-regulated learning (SRL) and explain why it is expected to prove particularly powerful in meta-cognitive tutoring system design.

Addressing (a) above, Sec. 2 defines the conceptual framework and the formalism of mental states, as they were developed and used in computational implementations of GMU-BICA[1] (the biologically-inspired self-aware cognitive architecture developed at George Mason University). Since a detailed description of GMU-BICA was never published, the present work aims at filling the gap from the top cognitive level by introducing the formalism of mental states and the basics of their dynamics.

Addressing (b) above, Sec. 3 is focused on a promising domain of potential applications of GMU-BICA: individual intelligent tutoring systems that can provide adaptive assistance to students in acquisition of self-regulation skills without interruption of the learning process. The formalism of mental states is a perfect fit for self-regulated learning (SRL) models. One of the expected outcomes, in addition to the impact on education, will be a blueprint of future intelligent artifacts possessing human-level learning skills. Before we start, we need to put this work in context.

1.1. *Background and motivation*

Despite tremendous progress and growing interest in artificial intelligence, neuroscience and cognitive science, a wide gap separates the approach of highly engineered top-down artificial solutions to narrow cognitive problems from the approach of bottom-up replication of robust, flexible and highly general solutions found in biological cognitive systems. In order to bridge the gap, it is essential that we better understand at a higher conceptual and computational level how biological systems naturally develop their cognitive and learning functions. In recent years, biologically inspired cognitive architectures (BICA) have emerged as a powerful new approach toward gaining this kind of understanding. Among them, GMU-BICA is focused on the higher-order, self-aware cognition. The architecture is “self-aware” because all cognitive representations in it are explicitly attributed to instances of the Self. This attribution is a fundamental aspect of human-like cognition.[2] As a result, GMU-BICA allows one to model human states of self-awareness, as explained below.

The general BICA challenge, as treated at the AAAI Fall Symposium on BICA,[3] is to create a computational equivalent of the human mind in its higher cognitive abilities. Arguably, it is vital for this purpose to use a biologically inspired approach, based on the principles of human self-aware cognition and learning. This approach will allow us to understand at a computational level how the machinery of the brain-mind develops the abilities to control attention, perceive objects and events, understand situations, reason about the past, plan for the future, learn from experience, and most importantly, as a result — grow cognitively from a baby to an adult. One of the ideas of a solution to the BICA challenge is to use a *cognitive chain reaction*: a process through which new cognitive functions and learning mechanisms emerge step by step from interactions and integration of existing ones, allowing an initially naïve artifact (yet possessing a *critical mass* of learning capabilities) to reach a human level of intelligence in selected domains via bootstrapped learning scaffolded by an instructor. This approach demands the level of robustness and flexibility of learning mechanisms that today is available only in biological systems. In particular, it requires the complex of higher-order mechanisms and strategies used by human learners that are labeled in the educational literature "self-regulated learning" (SRL)[4] and require meta-cognitive self-awareness.

These arguments bring us to the idea that a computational implementation and usage of human-like mental states in artifacts is a vital necessity for further progress in artificial intelligence, and may be a key to creating human-level intelligent artifacts. As a first step along this road, the underlying theoretical concept of a mental state needs to be specified at a computational level, and probably outside of the box of formal logic.[5,6] The laws of mental state dynamics need to be understood clearly before they can be implemented in artifacts. The first objective of the present work is to develop this kind of understanding of mental state dynamics by introducing a specific mental state formalism developed as part of the design of GMU-BICA.

1.2. *Brief overview of GMU-BICA*

GMU-BICA[1,7] includes eight highly interconnected biologically-inspired components (input-output, working, semantic, episodic and procedural memories, the neuromorphic cognitive map, the driving and reward-value systems — all mapped onto the brain[1]) that are essentially based on three building blocks: a *schema*, a *mental state*, and a *cognitive map*.[a] The cognitive map used in GMU-BICA is not considered here (see Ref. 8 for the cognitive map model).

All declarative representations (including input-output, working, semantic and episodic memories) are based on schemas. Schemas of GMU-BICA can represent any cognitive categories (from very abstract to very specific concepts) and are used similarly to object classes or LISP functions. For the present purposes, it is sufficient to accept that a schema is a universal elementary building block for all cognitive representations: e.g., a schema can be used to represent an object, an intended action, an observed relation, a goal situation, etc. Schemas may have names that capture their semantics, and this is how they will be represented here in mental state diagrams: by names. Their internal structure and mechanisms will be addressed elsewhere (some details are given in Refs. 7 and 9). Schemas bind to each other following the binding rules (that are also beyond the scope of this work), and through this binding, arbitrarily complex representations in principle can be formed from a limited set of initially given schemas. Thus, bound combinations of schemas can be learned and become new schemas. All available schemas constitute semantic memory of the agent; instances of schemas that are engaged in the ongoing information processing are present in working memory.

[a] These familiar terms are used here in a special sense and should not be confused with their other semantics.

While schemas are very important for understanding how the architecture works, the present focus is on the main top-level distinguishing feature of GMU-BICA, which is the representation and utilization of human-like mental states. Mental states are constructed from instances of schemas and populate working and episodic memories only. Mental states and their relations are dynamic elements of the architecture. They are created in working memory and upon deactivation are stored in episodic memory, from which they can be retrieved when necessary. Each case of retrieval and re-processing of a mental state may result in creation of a new copy of the mental state in episodic memory, which is consistent with the human data.[10]

2. Mental State Basics

2.1. *Concept*

Consider a cognitive architecture embedded in a virtual or physical environment. This notion implies that the agent (the architecture) is capable of perception, cognition and action, with elements of cognition (call them schemas) referring to features in the world, including the environment and the architecture. This general layout is characteristic of the majority of modern cognitive architectures, including recent versions of Soar[11] and ACT-R.[12]

The notion of a mental state that will be introduced here requires one step further beyond the aforementioned architectures. Specifically, it relies on the notion of a Self (an "I", a subject), which is very nontrivial in the context of computer science. In general, there are too many notions of a self used in the literature, and philosophers debate whether "a self" is philosophical nonsense (e.g., Refs. 13 and 14). None of this is relevant here. For the present practical purpose, the main property of a Self (or "I") is that a content of awareness can be attributed to it, and therefore this is how the notion of a Self (or "I") is understood below: a structureless, abstract token to which contents of mental states can be attributed, rather than the cognitive system itself or any of its observable aspects.[2] The key notion here is a *mental state*, understood as it was originally introduced in Ref. 2: "*…the entire complex involves two components: one representing the content of the experience of which the subject is aware, and the other labeling the subject who is experiencing this content* (*even though the subject may not be aware of self at the moment*). *Again, we call this complex a mental state*". Note that self-awareness of the subject is not required and may not be a part of a mental state content.

Relating to GMU-BICA, three details should be added. (i) An instance of a Self, or the "I" token, is always associated with a specific *mental perspective* of a subject, i.e., a view from a certain point in space, certain moment of time, certain agent's identity and status (present, imaginary, etc.), certain scale of distances, etc. One could say that a Self understood in this functionalist sense *is* the mental perspective of the subject. (ii) "I" tokens in working memory can be multiple, each representing a unique mental perspective. (iii) Contents of the subjective experience (i.e., awareness) attributed to "I" tokens consist of instances of schemas that are constrained by the self axioms[1,2] (not considered here).

Definition 2.1. A *mental state* in GMU-BICA is a set of instances of schemas attributed as the content of awareness to a unique mental perspective of a subject. Therefore, as a data structure, a mental state consists of two parts: the perspective of a subject and the content interpreted as a snapshot of awareness of that subject. For convenience, the *type* of the perspective is captured by a *mental state label* included in the data structure. Examples of mental state labels are: *I-Now*, *I-Next*, *I-Previous* (explained in Sec. 2.4). Labels are used in this work to refer to mental states themselves, even if there is ambiguity. All elements of mental states, including labels, are dynamic, change over time (e.g., *I-Now* typically becomes *I-Previous*), and obey the self axioms.[1,2]

The unique mental state that is required to be present in working memory at any time whenever the architecture is "awake" is labeled *I-Now*. It represents the current awareness of the agent about the current situation and has exclusive access to the input-output buffer: only *I-Now* can initiate actions or voluntarily shift attention of the agent (it is a scholastic question whether only the content of *I-Now* — or the entire content of working memory should be associated with "consciousness" of the agent). Other mental states are constructed by analogy with *I-Now*, and in many cases they either were at some point or can become *I-Now*, as explained below. The content of *I-Now* does not include all knowledge available to the agent in the current situation (called the epistemic, or knowledge state[15]). Indeed, psychological studies show that the span of human awareness is very limited;[16,17] therefore, in GMU-BICA, the number of instances of schemas that can be present in a mental state is limited, and so is the number of mental states in working memory. At the same time, consistent with biology, there is virtually no limit on the number of mental states stored in episodic memory.

2.2. *Structure*

The content of awareness represented by instances of schemas in *I-Now* can be structured into several overlapping domains, based on semantics and functionality. The *domain of control* includes elements that the agent can access and modify at any time. The *domain of perception* includes elements that are immediately available for sensory perception. The *central domain* may include the agent (labeled "me") and any facts that hold in the current context. Other domains may represent the notions of safety, desire, possibility, and so on. Structuring into domains provides a superficial, phenomenological description.

A more general approach consists of attaching a special attribute called *attitude*[1,7,9] to each instance of a schema (in the GMU-BICA framework, schemas have a number of standard attributes). The attitude of an instance of a schema specifies the status or position of the instance relative to the current mental perspective of the subject. Examples of *attitude components* are: "previous", "next", "intended", "desired", "10 feet ahead", etc. If the rule is that only those components of attitudes are specified explicitly that differ from the default values "actual", "present", "of myself", etc. (i.e., deviate from the center of the current mental perspective), then any instance in the central domain should have *nil* attitude. If attitude components can be viewed as coordinates in some abstract multidimensional space, then this *cognitive space* gives a structure to mental states. One of the main dimensions of the cognitive space is the *subjective time*: i.e., the time stamp associated with the instance. It generally differs from the physical time at which the instance is being processed.

Another aspect of the mental state structure relates to connections and interactions among instances of schemas. This topic is beyond the scope of the present consideration. It should be pointed out that in GMU-BICA schemas and their actions on each other are confined within each mental state, and the only allowed kinds of interaction of schemas across mental states are the acts of copying (or "projecting") schema instances from one mental state to another, with the appropriate modification of their attitudes, and subsequent "synchronization" of mirrored instances. Although the mirroring approach requires redundancy, it allows for an efficient partitioning of the global information processing in working memory of GMU-BICA into a set of restricted parallel processes, one per mental state.

2.3. *Types and functions*

Features of a given world can be described in various forms: e.g., using the event calculus,[6,18] using productions and operators of Soar, using propositions and predicates, using natural language, etc. Traditionally, representation of knowledge in artificial intelligence does not refer to any subjective perspective: it is supposed to represent objective facts rather than experiences. This is also true about representation of beliefs of agents: the statement "A believes that P" may not be attributed to any subject. In contrast, the mental-state-based approach always assumes a particular mental perspective of a subject, from within which the world is being perceived and understood. Therefore, in a sense, this approach is mind-centered rather than world-centered.

Another unique, distinguishing feature of GMU-BICA is the multiplicity of types of mental perspectives that can be represented simultaneously in its working memory by different mental states (Fig. 1). Multiple mental perspectives are used to simplify the processing of attitudes at different moments in time and at different stages of reasoning, while keeping the span of attention limited. The notion of the multiplicity of co-active mental states in working memory is biologically plausible and consistent with experimental and clinical data.[2,19] Interaction and dynamics of co-active mental states constitute the main mechanism underlying information processing in GMU-BICA.

There are at least two ways how a human subject can think about events and entities that do not take place in the current situation: (a) using objective theorizing,

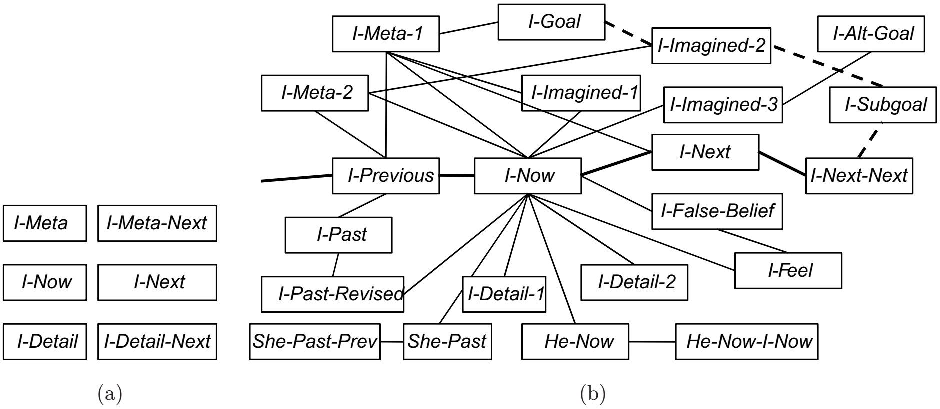

Fig. 1. (a) The central fragment of a two-dimensional slice of the mental state lattice. (b) Examples of possible mental state types that may be co-active in working memory (a possible mental state assembly). The fat line represents the working scenario: the sequence of mental states that were, are, or are expected to become *I-Now*. Its dashed part is not validated yet. Thin solid lines represent other essential dependence relations among mental states, e.g., subordination.

when the concept of an entity is brought into working memory with an understanding that it refers to another moment in time or to a virtual possibility, etc., and (b) using "mental time travel",[20] when the subject suspends her awareness of the current situation and imagines being in a different place and/or time, sometimes in a different world, or being somebody else, etc. In the mental state formalism, (a) corresponds to attitudes attached to instances of schemas, and (b) corresponds to the utilization of other mental states that represent different mental perspectives. Now it becomes clear that both attitudes and perspectives sample one and the same cognitive space, elements of which can be identified as possible attitudes, or as possible perspectives, or, speaking more generally, as possible generalized contexts (there is, however, a disagreement in the literature on the semantics of the word "context"[21,22]).

Mental states of different types (Fig. 1(b)) play different functional roles, and have different restrictions on their dynamics. The exclusive role of *I-Now* is explained in Sec. 2.1. Dynamics within *I-Now* are restricted by the schemas that capture common sense laws and general knowledge. *I-Past* is in addition constrained by the remembered facts. In contrast, certain kinds of *I-Imagined* allow for breaking the laws of common sense and suspending general beliefs about the world. Mental states like *He-Now* may allow for suspension of morality, etc. All mental state dynamics, however, are constrained by self axioms[1,2] that are hardcoded in the architecture (self axioms are not considered here). Interactions among mental states are determined by their types and contents. For example, *I-Next* inherits its content from *I-Now*, while *I-Meta* operates on a number of mental states and may actively modify their content. Each of its *subordinate* mental states is represented in *I-Meta* by an instance of the *agent schema* (e.g., the symbol "he" within *I-Now* in Fig. 3).

Consider a simple example (Fig. 2) that illustrates some of these rules. If a robot perceives a tree that is 10 feet ahead of it, then the attitude of the instance "tree" in

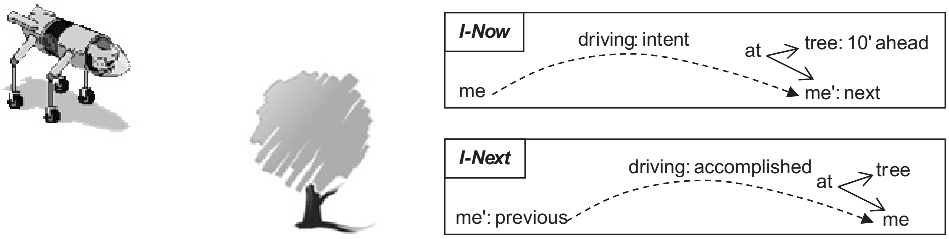

Fig. 2. An example of a situation (left) and the related mental state diagram (right). Both mental states *I-Now* and *I-Next* are simultaneously represented in working memory. Instances of schemas "me", "driving", "tree" and "at" are projected from one state to another with the appropriate attitude change. Upon reaching the tree, *I-Now* will become *I-Previous*, and *I-Next* will become *I-Now*.

I-Now is "10 feet ahead". If the robot intends to drive to the tree, then "driving to the tree" (an instance of the action schema in *I-Now*) has the attitude "intent". The imagined own body positioned next to the tree at the next moment of time labeled me′ ("me-prime") will have the attitude "next" in *I-Now*. The representation of the own body at its present position has the attitude *nil* in *I-Now*. At the same time, the robot has in its working memory a mental state *I-Next* reflecting expectations for the near-future experience. The content of *I-Next* will include an instance of the schema of driving with the attitude "accomplished".

2.4. *Mental state lattice*

If one can think of mental perspective (or attitude) components as discrete dimensions in the cognitive space (discrete spatial coordinates, discrete subjective time, discrete levels of meta-cognition, discrete levels of imagination, discrete Theory-of-Mind[23] levels, etc.), then the resultant multidimensional integer lattice in the cognitive space gives possible types of mental states. It is called the *mental state lattice* (practically it appears reasonable to limit possible mental state types to the integer lattice nodes; this does not imply that the system cannot live in, perceive, or reason about continuous space-time). Some of the generators of this lattice are: Next, Previous, Imaginary (or Imagined), Meta, He, She, I. Here is a simple algebra of these operators acting on the perspective of *I-Now*:

- Next (*I-Now*) = *I-Next* (this is my expected state of awareness at the next moment of time);
- Previous (*I-Now*) = *I-Previous* (this is my state of awareness at the previous moment of time, opposite to Next);
- Imaginary (*I-Now*) = *I-Imaginary* (this is a state of awareness of imaginary me);
- Meta (*I-Now*) = *I-Meta* (this is a state of awareness of me looking at my current episode of cognition from above);
- Detail (*I-Now*) = *I-Detail* (this is a state of awareness of me focused on one detail of my current awareness, ambiguous, opposite to Meta);
- He (*I-Now*) = *He-Now* (this is my simulation of his current state of awareness);

- I (*He-Now*) = *He-Now-I-Now* (this is my simulation of his belief about my state of awareness);
- She (*He-Now*) = *He-Now-She-Now* (this is my simulation of his current belief about her current state of awareness);
- Next (*I-Next*) = *I-Next-Next* (this is my expected state of awareness at time $T + 2$, if the current subjective time is T); etc.

In this list, most labels (e.g., Imaginary, Meta, Detail) ambiguously specify operators and their actions, and may have many flavors. For example, Imaginary could be Hypothesized or Assumed, plus the imagined content may differ from case to case; Meta could differ by the kind of meta-cognitive perspective (e.g., self-evaluation based on standards or based on feelings), etc. These details need to be specified for each particular computational implementation.

Each attitude of a schema instance that is present in a mental state refers to a lattice node. Therefore, attitudes suggest creation of new mental states (if those states are missing) so the reasoning process can be continued. Not always those mental states will be created by the architecture, because of the limit on the number of mental states in working memory. Mental states that are actually present in working memory at a given moment of time occupy a small subset of the lattice nodes. They constitute the active *mental state assembly* associated with one currently experienced *episode.* In episodic memory, mental states are stored together with their relations and references to each other, as one global network. An episode corresponds to a fragment of this network that can be retrieved into working memory. In GMU-BICA, attitudes are represented in an egocentric (i.e., subject-centered) reference frame, relative to the current mental perspective of the subject. For perspectives, both egocentric and allocentric (subject-independent and permanent) labeling is used. Throughout this work, egocentric mental state labels that are defined relative to *I-Now* are used in diagrams.

2.5. *Dynamics*

The contents of mental states in Fig. 2 change with time: the attitude of "driving" in *I-Now* becomes *nil* when the robot starts driving. The attitude of "tree" becomes "9 feet ahead", "8 feet ahead", and so on, as the robot approaches the tree. Finally, when the tree is reached, *I-Now*, instead of changing its content, changes its label to *I-Previous* (subsequently this *I-Previous* will be stored in episodic memory with its permanent allocentric label, then, when necessary, this stored mental state could be retrieved back to episodic memory with the label *I-Past*, and the sequence of the events of driving could be replayed). At the same time, *I-Next* is validated by the current sensory input. If it passes the validation, then *I-Next* becomes *I-Now.* Mental states change labels whenever something significant happens or the limit of time allocated for them expires: each mental state corresponds to one elementary episode (although an entire episode is typically represented by a small transient assembly of mental states: see above).

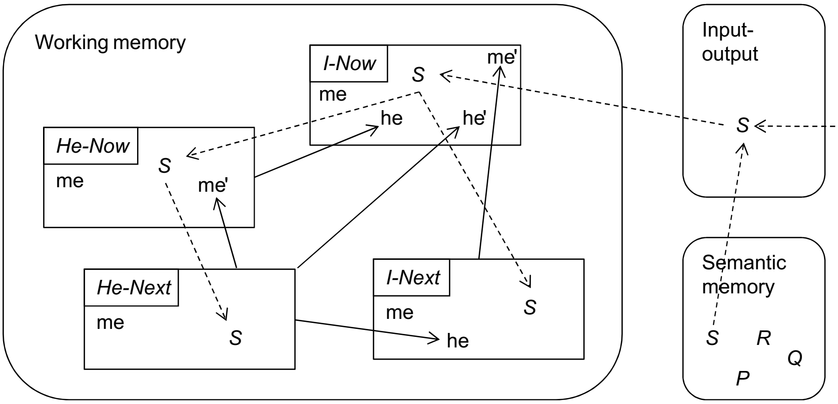

Fig. 3. Sensory perception with shared attention. Solid arrows represent attribution of mental states (rectangular boxes) to instances of the agent schema (me, he, me′, he′). Schemas P, Q, R, S are stored in semantic memory. Dashed arrows represent the steps of copying of instances of the schema S into all mental states where it should appear. Mental state labels are typed in italic in the upper left corner of each mental state box.

Sensory perception with shared attention (Fig. 3) is another example illustrating perception of sensory input, Theory of Mind[23] and interactions of mental states. Suppose the agent is aware of another agent present in the same environment. That agent is represented in working memory by a mental state *He-Now* and by an instance "he" of the agent schema in *I-Now*, to which *He-Now* is attributed. An incoming sensory signal is represented by an instance of schema S (Fig. 3) in the input-output buffer, from which the filtered by attention content is copied into *I-Now*, and from there to *He-Now*, and (if the perceived entity is expected to persist at the next moment of time) to *I-Next* and to *He-Next*. Now we see that mental states can suggest instances of schemas, and vice versa.

In a more general case represented by the mental state diagram of Fig. 1(b), many mental perspectives are processed in parallel. The role of *I-Now* is sequentially performed by *I-Previous*, *I-Now*, *I-Next*, *I-Next-Next*, etc., if everything goes as expected. This expected sequence of *I-Now*'s is called the *working scenario* (a working scenario has to be a non-contradictory ordered sequence of mental states connecting *I-Now* and *I-Goal*). However, deviations from the working scenario are allowed and may become necessary. For example, if at a certain point the agent should focus and consider separately a small detail of the plan, then, e.g., *I-Detail-2* should temporarily become *I-Now* (Fig. 1(b)), and for the duration of this episode, the original *I-Now* becomes *I-Meta*. The reason for having *I-Detail-2* in working memory at the time when everything goes as expected is to get ready for the possible challenge, using parallel background resources. As a result, the agent can almost momentarily switch from using one perspective to another in its behavior, while continuing working on other perspectives in the background.

Similarly, it might become necessary to evaluate the entire progress from a certain meta-cognitive perspective: in this case, e.g., *I-Meta-1* becomes *I-Now* (while *I-Now* becomes *I-Detail*, and *I-Goal* may be temporarily working as *I-Next*). Another possibility is a conflict between expectations and the outcome: e.g., if *I-Next* does not pass validation by the sensory input, then it becomes *I-False-Belief*, a new *I-Now* is created based on sensory input, and the agent may consider pursuing a previously rejected route: e.g., via *I-Imagined-1*, or jumping to *I-Meta* and trying to understand what happened at a higher level, then possibly revise semantic knowledge and/or the current goals.

These are merely a few examples of possible behaviors. Typically, the behavior of GMU-BICA consists of cycles of *voluntary action*. The cycle includes the following phases (with possible branching at any point, which is not considered here).

- Perception and understanding of the situation, validation of the expectations.
- Generation of ideas of feasible and reasonable actions that can be done immediately.
- Intent generation by selection and validation of ideas that fit into working scenario.
- Execution of intended and scheduled actions (each action is not necessarily physical: it could be a cognitive action, e.g., making a commitment or creating a new schema).

Behaviors like self-analysis or meta-cognitive reasoning during learning may be based on different patterns, yet in principle they are reducible to sequences of voluntary acts.

2.6. *Section summary*

In this section, we introduced the basic formalism of mental states of GMU-BICA by specifying the important conceptual, structural and dynamical aspects of mental states, their parts and assemblies, viewed as dynamical systems populating working and episodic memories of the cognitive architecture. Many important details, even at this top level, escaped our consideration. A complete, detailed presentation would require a long specification of other levels and components of GMU-BICA as well, primarily including the formalism of schemas. Nevertheless, the vital minimum presented above should allow the reader to understanding the following section, which lays out theoretical grounds for a highly promising potential application of GMU-BICA: in meta-cognitive tutoring systems. Our next task is to explain the feasibility and the power of application of the presented mental-state-based approach to modeling higher cognitive and learning processes that engage self-regulation and meta-cognition.

3. Mental-State Model of Self-Regulated Learning (SRL)

3.1. *Connecting mental state formalism to SRL*

Approaches based on self-regulation and on intelligent tutoring systems nowadays are rapidly acquiring popularity in educational practice and merge together. Here we explain why developing an SRL tutoring assistant is an ideal application for a general-purpose BICA based on the mental state formalism introduced above, by connecting the two models step-by-step.

SRL involves deliberative construction of knowledge by using goals, strategies, self-monitoring and self-evaluation. Self-regulated learners are meta-cognitively, motivationally and behaviorally active participants of their own learning.[24] It is interesting to notice that general models designed for SRL in educational literature[4] come very close to mental state diagrams like those in Fig. 1 and can be implemented on the basis of mental states. Computer-based SRL models and assistants used in education today do not perform simulations of the student awareness and self-awareness using cognitive architectures. A good SRL assistant will implement a dynamic, adaptive functional model of the SRL process based on a cognitive architecture.

Figure 4 represents an attempt to translate a traditional SRL model[4,25] into a mental state diagram compatible with GMU-BICA. The SRL model has three phases: forethought, performance, and self-reflection, that are cyclically interrelated in a self-oriented system of feedback. The main feedback loop (Fig. 4, top) resembles the voluntary action cycle explained in Sec. 2.5, although it works at a larger time scale.

The *forethought phase* precedes action and involves (a) an analysis of the task to be regulated, including setting goals and selecting strategies to accomplish the goals,

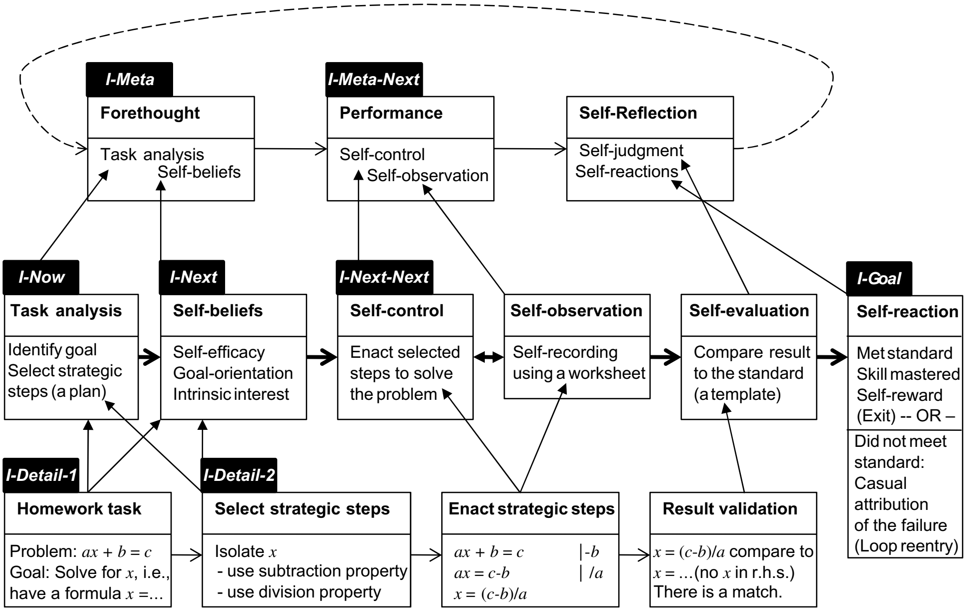

Fig. 4. SRL model (based on Ref. 4) underlying the process of solving Problem 3.1 formatted to fit the mental state formalism. SRL labels are typed in bold, mental state labels (showing a snapshot of working memory) appear on the black background. All mental states that are labeled in the diagram are active at the represented moment of time close to the beginning of the solution process. *I-Goal* is associated with the upper half of the box labeled "Self-reaction". Horizontal arrows indicate the sequence of phases, and vertical arrows show subordination of (potential) mental states. The working scenario is represented by fat arrows.

and (b) student motivation consisting of self-beliefs such as self-efficacy, intrinsic interest in the task, and mastery goal orientation. In terms of GMU-BICA, (a) implies instantiation and binding of schemas of strategies and subgoals, while (b) involves setting attitudes "can do", "interested in doing", etc., and corresponding other parameters of those instances. Students oriented toward a learning goal are more likely to self-regulate than students with a performance goal orientation. This property can follow directly from GMU-BICA dynamics, if learning-goal orientation creates a bias toward higher meta-cognitive levels.

The next phase, the *performance phase*, consists of two categories: self-control and self-observation. The learner is now focused on the task implementing his/her plan of action and monitoring progress by using procedures such as self-recording, graphing, logbooks, and journals. Here, GMU-BICA offers tools for recording and analysis of its own dynamics, from episodic memory to meta-cognitive introspection. During the self-control phase, learners use strategies such as self-instruction, which is a useful method of reminding them to use appropriate strategies. Again, this element can be conveniently implemented in GMU-BICA using meta-cognitive perspectives for self-control.

During the final phase of the model, the *self-reflection phase*, learners engage in self-evaluation by analyzing self-monitored outcomes to judge their performance against a standard. Therefore, it is in this phase where learners make attributions to potential causes of outcomes and decide whether the outcomes are optimal or need improvement.

In the latter case, learners would attempt to re-adjust their strategic plan by selecting new strategies through the forethought phase and attempt to perform the task again. This self-regulation cycle can be used by novice learners who are simply learning a new task, or by experts who are trying to improve/eliminate unforced errors during performance efforts. All these elements can be naturally reflected in GMU-BICA, as we explain below.

In the next subsection we consider a specific SRL model with a mental state diagram mapped onto it, so that it can be used to implement a computational model of SRL as a functional component in an intelligent SRL assistant based on GMU-BICA.

3.2. *Designing a computational model of SRL based on mental states*

To make a better connection between SRL and the mental state framework at a computational level, we consider an example of a problem solving paradigm and construct its description that unifies both frameworks: the SRL model and the mental state formalism. For this purpose, we select the following homework assignment in high-school algebra.

Problem 3.1. Solve for x:

$$ax + b = c, \tag{3.1}$$

$$ax^2 + bx + c = d. \tag{3.2}$$

Facts and rules known to the student at this point include the multiplication, division, addition and subtraction properties of equations, the quadratic formula, etc.

The SRL model that may be used in this case by a self-regulated learner is shown in Fig. 4 in a form that suggests its implementation using GMU-BICA. The diagram of Fig. 4 represents SRL phases of problem solving as boxes organized into three levels: the task level (applied to Problem 3.1), the first meta-cognitive level, and the second meta-cognitive level. Horizontal arrows indicate the sequence of phases, and vertical arrows show subordination of (potential) mental states: each box can be associated with a mental state of the learner who is working on this phase of problem solving. We see that in this case the network of SRL phases (represented by connected boxes in Fig. 4) plays the role of a mental state lattice described in Sec. 2.4.

From the GMU-BICA point of view, the task of implementing SRL based on the mental state formalism is to map boxes onto mental states and to instantiate the mental states in working memory of the architecture, to map the content of each box onto schemas representing elements of SRL (e.g., strategies) and to establish connections between instances. When this is done, and the rules of schema binding correctly reflect relations among elements of SRL and the curriculum, the resulting system can be used to simulate the process of student learning. A simulation of this sort can be incorporated as a cognitive component into a tutoring system in order to diagnose problems that a student may experience and provide the right scaffolding at the right time, or to demonstrate to an SRL-novice student practically how the steps of SRL should be performed.

When the solution method represented in the bottom row in Fig. 4 is applied to the second problem (3.2), it fails (not shown in Fig. 4): the strategy of isolating x using the subtraction and division properties of equations is not effective in this case. Trying to isolate x in a quadratic equation is a common mistake among high-school students. The system (Fig. 4) should be able to deal with the failure successfully, as follows. At the first step after the failure is detected, *I-Now* takes the position of the lower half of the box labeled "Self-reaction". At this step, attribution of the cause of the failure is made to a selection of inefficient strategies. Therefore, the agent starts the loop over, with *I-Now* returning to "Task analysis" (Fig. 4, left), now paying more attention to selection of strategies and rejecting the original choice. At the same time, *I-Next* associated with the "Self-beliefs" box maintains its belief in self-efficacy, its mastery-goal-orientation, and attributes the "can do" attitude to the steps considered at the task level, despite the recent failure. This results in an overall positive attitude of the agent toward the problem. Under these circumstances, the agent is likely to find and use the right strategy (which is to apply the quadratic formula to Problem 3.2 after an elementary transformation), and creates a new plan of solution, which is then executed. Here positive motivational beliefs are very important and cannot be compensated by knowledge of how and when to use strategies. The learner should be interested in the task and should have positive beliefs of her own capabilities in order to master the problem solving skill.

In conclusion, we have demonstrated that the diagram (Fig. 4) can be used to build a computational model of SRL. The relative frequency of usage of the higher layers of SRL phases is a parameter of implementation determined by attitudes and beliefs. It needs to be optimized in order to achieve better performance in learning and in order to match student data. This can be done using computational and classroom experiments.

4. Discussion

4.1. *Related approaches and implications for SRL tutoring*

All the best-known cognitive architectures are "biologically-inspired", in the sense that they were inspired by studies of natural cognition and/or the brain, and therefore capture principles of operation of the human mind and/or functional neuroanatomy of the central nervous system. Examples are well-known cognitive architectures such as Soar,[11,26,27] ACT-R,[12,28] LEABRA,[29] SAL,[30] EPIC,[31] TOSCA,[32] CLARION,[33] LIDA,[34] Polyscheme,[35] plus their integrations and modifications.[3] None of these architectures, however, can be confidently placed at the highest level of the widely known hierarchy of intelligent agent architectures summarized in Table 1.

In contrast, GMU-BICA by its design belongs to the highest level in Table 1, as explained above. The most popular cognitive architectures (e.g., the extended Soar[27]) are evolving toward the top level of Table 1, acquiring mental-state-like features that allow them to implement "episodic memories", yet still are missing many essential features. Among related practically useful approaches based on mathematical logic, the event calculus[6,18] is probably the closest formalism to schemas and mental states of GMU-BICA, and yet it does not include tools for modeling mental states at the same level. New modal logics try to cope with the phenomenology of mental states.[36] Terms "mental state", "attitude" and the like have been used in various versions of mathematical logic at least for several years,[37,38] yet rigorous formal logic requirements complicate the progress in their practical applications. It is also clear that the brain operates differently.

Table 1. Hierarchy of intelligent agent architectures.

Cognitive architecture type	The agent is capable of	Level
Meta-cognitive and self-aware	Modeling mental states of agents, including own mental states, based on the self concept	Highest
Reflective	Modeling internally the environment and behavior of entities in it	High
Proactive, or deliberative	Reasoning, planning, exploration and decision making	Middle
Reactive, or adaptive	Sub-cognitive forms of learning and adaptation	Low
Reflexive	Pre-programmed behavioral responses	Lowest

The topic of meta-cognition acquires increasingly higher weight in all fields of interdisciplinary cognitive sciences, from artificial intelligence to education. An example is a recently developed general concept of a meta-cognitive architecture (developed by Cox and Raja[39]) that introduces multiple levels and multiple kinds of meta-cognition (or "metareasoning"). A unification of meta-cognition and SRL at a cognitive architecture level is highly desirable for education. However, existing implementations of SRL models in computer-based intelligent tutoring and diagnostic systems follow a different path, using statistical models rather than cognitive architectures, or cognitive architectures from one of the lower levels of Table 1. By providing limited, passive or reflexive SRL support to students, they essentially leave the problem open. An example is the Betty Brain tutoring system,[40] which uses the paradigm of tutoring by teaching and represents the world's state of the art in the field of intelligent tutoring. Its new version[41] employs a hidden Markov model (a statistical model, not a cognitive architecture) to describe transitions between SRL phases. Arguably, this approach is not sufficient for building a faithful model of human SRL that can be transferred to students, and it could benefit from adopting features of the mental-state model of SRL described here.

4.2. *Psychological underpinnings and implications*

One essential hypothesis in cognitive psychology inspiring the mental state formalism is that there are multiple, simultaneously active and interacting with each other mental state representations in human working memory, each associated with its own subjective perspective. This counterintuitive hypothesis is supported by empirical and clinical data. For example, studies of the human Theory-of-Mind phenomenon suggest that simulations of other minds occur in the brain automatically and in parallel with processing of the subject's own state of awareness.[42] These observations are consistent with the widely accepted today simulationist account of the human Theory-of-Mind,[43] which is also supported by many empirical brain data, such as studies of mirror neurons in humans.[44,45] A comprehensive review of supporting evidence is given in Ref. 2, which provides the main background for the presented framework. Another foundation is the well-known Tulving's theory of episodic memory.[46,47] In general, psychological and neurological literature underlying and supporting the multiple mental state concept is immense, and cannot be properly reviewed here. To give one more example, Ericsson and Kintsch[48] introduce an empirically supported concept of "long-term working memory", implying that a number of working memory states, each representing its own context or mental perspective (using our terminology), can be swapped back and forth between working and long-term memory during one and the same reasoning episode at the speed of normal working memory operations. This view supported by experimental data seems hard to reconcile with current models of long-term memory retrieval (including the one used for GMU-BICA design[8]). It leads, on the contrary, to expansion of working memory models.[49] Consistent with this outcome, the multiple mental state model offers an explanation of the data based on multiple mental states in working memory, predicting that the timing of reasoning involving multiple mental perspectives could be much faster than the timing of long-term memory access, if representations of other mental perspectives are processed in working memory in parallel.

4.3. *Concluding remarks*

In this work we presented a general theoretic formalism that allows us to build intelligent cognitive systems of the next generation, directly simulating features of operation of the human mind, including its learning abilities. Our previous prototypes[7] based on this formalism demonstrated its consistency and usefulness for practical applications. Today GMU-BICA continues to be a highly valuable prototype for intelligent artifacts in many domains, including intelligent tutoring systems. The present theoretical work focused on the top cognitive level of GMU-BICA: mental states, leaving detailed description of schemas and non-declarative components for future publications.

Among the distinguishing features of the GMU-BICA mental state formalism are: (i) a subject-centered view of the world, (ii) the multiplicity of mental perspectives simultaneously represented in working memory, each playing its unique functional role, and (iii) the limited span of awareness. All these features are consistent with human psychology. The present work supports and exemplifies a view that these biological constraints are vital for flexibility, robustness and effectiveness of artifacts deployed in new/unexpected domains, such as SRL modeling, one outcome of which will be a blueprint of future artifacts possessing human-level learning capabilities.

Finally, in this work we demonstrated that the constructed theoretical framework can be used to build computational models of SRL for pedagogical agents. We illustrated the power of the approach based on mental state formalism for the design of computer-based SRL assistants. A personal assistant based on GMU-BICA will be able to provide interactive help to student learners at a higher cognitive and metacognitive level, while at the same time being able to adjust itself dynamically to each student individually. This particularly valuable combination of features is missing in today's intelligent tutoring systems.

Acknowledgments

We are grateful to Dr. Carol L. O'Donnell, the US DOE IES NCER CASL Program Officer who encouraged us to use BICA models of meta-cognition and self-awareness for intelligent SRL assistants. The presented here formalism of mental states was inspired by early discussions in Tucson, AZ, with Drs. Lynn Nadel, Alvin I. Goldman, and others. We are grateful to Dr. Nada Dabbagh, an Associate Professor of Instructional Design and Technology in the college of Education and Human Development at GMU and a member of our team, who helped us to connect the BICA framework to intelligent tutoring systems. We are also grateful to Ms. Laurie Lewis,

an 8th Grade teacher at Arlington Public Schools in Virginia, who provided us with many useful examples of mathematical problems and difficulties that students have with them.

References

1. A. V. Samsonovich and K. A. De Jong, Designing a self-aware neuromorphic hybrid, in *AAAI'05 Workshop on Modular Construction of Human-Like Intelligence: AAAI Technical Report* WS-05-08, eds. K. R. Thorisson, H. Vilhjalmsson and S. Marsela (AAAI Press, Menlo Park, CA, 2005), pp. 71–78.
2. A. V. Samsonovich and L. Nadel, Fundamental principles and mechanisms of the conscious self, *Cortex* **41** (2005) 669–689.
3. *Biologically Inspired Cognitive Architectures: Papers from the AAAI Fall Symposium. AAAI Technical Report* FS-08-04, ed. A. V. Samsonovich (AAAI Press, Menlo Park, CA, 2008).
4. B. J. Zimmerman, Attaining self-regulation: A social cognitive perspective, in *Handbook of Self-Regulation*, eds. M. Boekaerts, P. R. Pintrich and M. Zeidner (Academic Press, San Diego, CA, 2000), pp. 13–39.
5. M. Minsky, *The Society of Mind* (New York, Simon & Schuster, 1985).
6. E. T. Mueller, *Commonsense Reasoning* (Elsevier, Amsterdam, 2006).
7. A. V. Samsonovich, G. A. Ascoli, K. A. De Jong and M. A. Coletti, Integrated hybrid cognitive architecture for a virtual roboscout, in *Cognitive Robotics: Papers from the AAAI Workshop, AAAI Technical Report* WS-06-03, eds. M. Beetz, K. Rajan, M. Thielscher and R. B. Rusu (AAAI Press, Menlo Park, CA, 2006), pp. 129–134.
8. A. V. Samsonovich and G. A. Ascoli, A simple neural network model of the hippocampus suggesting its pathfinding role in episodic memory retrieval, *Learn. Memory* **12** (2005) 193–208.
9. A. V. Samsonovich, Biologically inspired cognitive architecture for socially competent agents, in *Cognitive Modeling and Agent-Based Social Simulation: Papers from the AAAI Workshop, AAAI Technical Report* WS-06-02, eds. M. A. Upal and R. Sun (AAAI Press, Menlo Park, CA, 2006), pp. 36–48.
10. L. Nadel, A. Samsonovich, L. Ryan and M. Moscovitch, Multiple trace theory of human memory: Computational, neuroimaging, and neuropsychological results, *Hippocampus* **10** (2000) 352–368.
11. J. E. Laird and P. S. Rosenbloom, The evolution of Soar cognitive architecture, in *Mind Matters: A Tribute to Allen Newell*, eds. D. M. Steier and T. M. Mitchell (Mahwah, NJ: Erlbaum, 1996), pp. 1–50.
12. J. R. Anderson and C. Lebiere, *The Atomic Components of Thought* (Mahwah, NJ: Erlbaum, 1998).
13. A. Sloman, "The Self": A bogus concept (2008), `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/the-self.html`
14. A. V. Samsonovich and G. A. Ascoli, The conscious self: Ontology, epistemology and the mirror quest, *Cortex* **41** (2005) 621–636.
15. J.-P. Doignon and J.-C. Falmagne, *Knowledge Spaces* (Springer, Berlin, 1999).
16. G. A. Miller, The magic number seven plus or minus two: Some limits on capacity for processing information, *Psychol. Rev.* **63** (1956) 81–97.
17. N. Cowan, The magical number 4 in short-term memory: A reconsideration of mental storage capacity, *Behav. Brain Sci.* **24** (2001) 87.
18. E. T. Mueller, Event calculus and temporal action logics compared, *Artif. Intell.* **170** (2006) 1017–1029.

19. A. V. Samsonovich, Hallucinating objects versus hallucinating subjects, *Behav. Brain Sci.* **28** (2005) 772–773.
20. M. A. Wheeler, D. T. Stuss and E. Tulving, Toward a theory of episodic memory: The frontal lobes and autonoetic consciousness, *Psychol. Bull.* **121** (1997) 331–354.
21. J. McCarthy and S. Buvac, Formalizing contexts (expanded notes) (1998), `http://www-formal.stanford.edu/jmc/mccarthy-buvac-98/context/context.html`
22. R. C. Stalnaker, *Context and Content: Essays on Intentionality in Speech and Thought*, Oxford Cognitive Science Series (Oxford University Press, Oxford, 1999).
23. S. Baron-Cohen, *Mindblindness: An Essay on Autism and Theory of Mind* (MIT Press, Cambridge, MA, 1995).
24. B. J. Zimmerman, A social cognitive view of self-regulated academic learning, *J. Educational Psychol.* **81** (1989) 329–339.
25. B. J. Zimmerman and A. Kitsantas, The hidden dimension of personal competence: Self-regulated learning and practice, in *Handbook of Competence and Motivation*, eds. A. J. Elliot and C. S. Dweck (Guilford Press, New York, 2008), pp. 509–526.
26. J. E. Laird, P. S. Rosenbloom and A. Newell, *Universal Subgoaling and Chunking: The Automatic Generation and Learning of Goal Hierarchies* (Kluwer, Boston, 1986).
27. J. E. Laird, Extending the Soar cognitive architecture, in *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, eds. P. Wang, B. Goertzel and S. Franklin (IOS Press, Amsterdam, 2008), pp. 224–235.
28. J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere and Y. Qin, An integrated theory of the mind, *Psychol. Rev.* **111** (2004) 1036–1060.
29. R.-C. O'Reilly and Y. Munakata, *Computational Exploration in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain* (MIT Press, Cambridge, 2000).
30. C. Lebiere, R. O'Reilly, D. J. Jilk, N. Taatgen and J. R. Anderson, The SAL integrated cognitive architecture, in *Biologically Inspired Cognitive Architectures: Papers from the AAAI Fall Symposium*, AAAI Technical Report FS-08-04, ed. A. V. Samsonovich (AAAI Press, Menlo Park, CA, 2008), pp. 98–104.
31. D. E. Meyer and D. E. Kieras, A computational theory of executive cognitive processes and multiple task performance: Part I. Basic mechanisms, *Psychol. Rev.* **63** (1997) 81–97.
32. J. E. Laird *et al.*, TOSCA: A comprehensive brain-based cognitive architecture (2006), `http://www.darpa.mil/ipto/programs/bica/docs/TOSCA.pdf.`
33. R. Sun, The CLARION cognitive architecture: Extending cognitive modeling to social simulation, in *Cognition and Multi-Agent Interaction*, ed. R. Sun (Cambridge University Press, New York, 2004).
34. S. Franklin, A foundational architecture for artificial general intelligence, in *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms, Proc. of the AGI Workshop 2006*, eds. B. Goertzel and P. Wang, Frontiers in Artificial Intelligence and Applications, Vol. 157 (IOS Press, Amsterdam, The Netherlands, 2007), pp. 36–54.
35. N. L. Cassimatis, J. G. Trafton, M. D. Bugajska and A. C. Schultz, Integrating cognition, perception and action through mental simulation in robots, *J. Robotics and Autonomous Systems* **49** (2004) 13–23.
36. E. Lorini and A. Herzig, A logic of intention and attempt, *Synthese* **163** (2008) 45–77.
37. P. Panzarasa, N. R. Jennings and T. J. Norman, Formalizing collaborative decision making and practical reasoning in multi-agent systems, *J. Logic Comput.* **12** (2002) 55–117.
38. A. F. Dragoni, P. Giorgini and L. Serafini, Mental states recognition from communication, *J. Logic Comput.* **12** (2002) 119–136.
39. M. T. Cox and A. Raja, *Metareasoning: A Manifesto*, Technical Report BBN TM-2028 (BBN Technologies, 2007), `http://www.mcox.org/Metareasoning/Manifesto.`

40. G. Biswas, K. Leelawong, D. Schwartz and N. Vye, Learning by teaching: A new agent paradigm for educational software, *Appl. Artif. Intell.* **19** (2005) 363−392.
41. H. Jeong, A. Gupta, R. Roscoe, J. Wagster, G. Biswas and D. Schwartz, Using hidden Markov models to characterize student behavior patterns in computer-based learning-by-teaching environments, in *Intelligent Tutoring Systems: 9th Int. Conf. ITS-2008*, Lecture Notes in Computer Science, Vol. 5091, eds. B. Woolf *et al.* (Springer, Berlin, 2008), pp. 614−625.
42. T. P. German, J. L. Niehaus, M. P. Roarty, B. Giesbrecht and M. B. Miller, Neural correlates of detecting pretense: Automatic engagement of the intentional stance under covert conditions, *J. Cognitive Neurosci.* **16** (2004) 1805−1817.
43. S. Nichols and S. Stich, *Mindreading: An Intergrated Account of Pretence, Self-Awareness, and Understanding Other Minds* (Oxford University Press, Oxford, 2003).
44. G. Rizzolatti and C. Sinigaglia, *Mirrors in the Brain: How Our Minds Share Actions and Emotions*, trans. Frances Anderson (Oxford University Press, Oxford, 2008).
45. V. Gallese and A. Goldman, Mirror neurons and the simulation theory of mind-reading, *Trends Cogn. Sci.* **2** (1998) 493−501.
46. E. Tulving, *Elements of Episodic Memory* (Oxford University Press, Oxford, 1983).
47. E. Tulving, Episodic memory: From mind to brain, *Rev. Neurol.-France* **160** (2004) S9−S23.
48. K. A. Ericsson and W. Kintsch, Long-term working-memory, *Psychol. Rev.* **102** (1995) 211−245.
49. I. R. Olson, K. Page, K. S. Moore, A. Chatterjee and M. Verfaellie, Working memory for conjunctions relies on the medial temporal lobe, *J. Neurosci.* **26** (2006) 4596−4601.