Self-Awareness as Metacognition about Own Self Concept

Alexei V. Samsonovich[1], Anastasia Kitsantas[2], Nada Dabbagh[2] and Kenneth A. De Jong[1,3]

[1]Krasnow Institute for Advanced Study, George Mason University, Fairfax, VA 22030
[2]College of Education and Human Development, George Mason University, Fairfax, VA 22030
[3]Computer Science Department, George Mason University, Fairfax, VA 22030
asamsono@gmu.edu, kdejong@gmu.edu, akitsant@gmu.edu

Abstract

Implementation of agency in a cognitive system implies that certain beliefs, values and/or goals represented in the system become, if implicitly, attributed to the self of the agent. When the cognitive system becomes explicitly aware of this attribution, it acquires a self-regulation capacity allowing it to control, modify and develop its self-concept together with the attitudes attributed to the self, adjusting to dynamically changing contexts and personal experience. The leverage of self-awareness understood in this sense consists in increased robustness, flexibility and integrity of the cognitive system, as illustrated by a paradigm of self-regulated learning.

Introduction

The topic of metacognition acquires increasingly higher weight in all fields of interdisciplinary cognitive sciences, from artificial intelligence (AI) to education. The general concept of metacognition (or "metareasoning", which is understood more narrowly: Cox and Raja 2008) is captured by Figure 1. It involves at least two levels of cognitive representations in the system: "object" and "meta" levels.

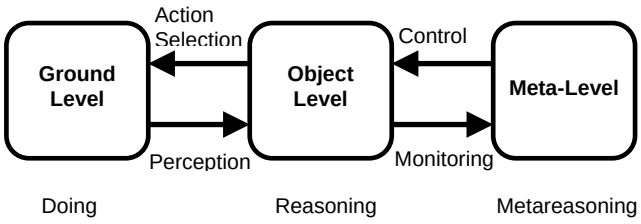

Figure 1. The general framework for metacognition in a cognitive agent architecture (from Cox and Raja 2008).

Taken merely as an architectural or syntactic constraint, this general functional scheme (Figure 1) in and by itself does not tell us what (if any) new cognitive quality will be introduced into the system with the addition of a metacognitive level. There are several interpretations of this scheme that give slightly different answers to this question (e.g., Russel and Welfad 1991, Cox and Ram 1999, Raja and Lesser 2007). The most interesting interpretation, which is the topic of the present work, relates to the concept of a self understood in one particular sense (Samsonovich and Nadel 2005, cf. Singh 2005). In order to put this topic into the general perspective of cognitive architecture design, it is useful to consider the following hierarchy of cognitive agent architectures, in which all six levels, except for the top one ("self-aware"), are generally well-accepted by the AI community:

- Reflexive (based on a fixed set of behavioral responses)
- Reactive (capable of learning and adaptation)
- Proactive, or deliberative (capable of reasoning, planning, exploration and decision making)
- Reflective (capable of modeling the environment and behavior, using mental simulations as part of cognition)
- Metacognitive (capable of higher-order representation and control of cognitive states of agents, including itself)
- Self-aware (see below)

Today there is no general consensus regarding the understanding of the term "self-aware agent". The notion of self-awareness involves the notion of a self, and that, in turn, has multiple semantics in the cognitive science and computer science literature, ranging from "self as own body" to "self as own identity". The specific notion of a *self* that is used here was introduced by Samsonovich and Nadel (2005) as a structureless abstraction to which all cognitive (mental) states in the system are attributed by the system itself. According to their work, this attribution is a "fundamental mistake" of the system itself, which together with the *self axioms* (that are built into the architecture) constrain cognition in the system, making its behavior look as if it was orchestrated by a unique "self". At that time, the leverage of having this self concept in a cognitive architecture was not satisfactory articulated (see, however, Samsonovich and De Jong 2005). The objective of the present work is to address the pragmatic side of the above self concept through analysis of the nature of mental states of self-awareness and their associated functional roles.

General Consideration of Self-Awareness

The difference between metacognition with and without self-awareness understood in the above sense can be explained as follows. Consider a cognitive architecture that

may have representations of various attitudes: beliefs, values, goals, intentions, etc., but lacks the concept of self, in the sense that there is no explicitly represented attribution of attitudes to any agent, if other agents are not involved (Figure 2 a). Nonetheless, one can formally interpret these attitudes as being attributed to the self of this cognitive system (the agent), and not others. The structure of this attribution, however, cannot be altered dynamically. For example, if the agent learns that some of its beliefs are false, then it can change the content of those beliefs, or label them as "false", etc. In either case, the cognitive states remain (implicitly) attributed to the same self. If the agent should alter its system of beliefs, goals, etc., then, generally speaking, it should revise its entire system of representations. This limitation remains in effect at the metacognitive level, where object-level beliefs are reflected, or new beliefs about the system and its abilities become instantiated. There is no room for the dynamic attribution of mental states in this framework (Figure 2 a).

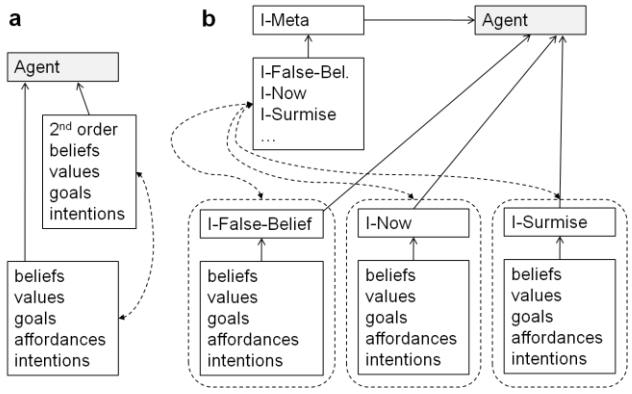

Figure 2. Metacognition (a) without and (b) with self-awareness. Solid arrows show attribution, dashed arrows show essential interaction. In (a), the attribution to the Agent is not explicitly represented in the system.

A different scenario becomes possible in a cognitive architecture where the current instance of the self (associated with the current time, place, status, etc., as well as with the identity of the agent) is represented explicitly by a token, and the attribution of mental states[1] to it is also represented explicitly (Figure 2 b; the tokens are self-explanatory mental state labels: "I-Now", "I-Meta", etc.: Samsonovich and De Jong 2005). In this case, multiple instances of a self with their own mental states may coexist in working memory (where they actively interact with each other), and also in *episodic memory* (Tulving 1983). For example, the agent may become aware of other agents by representing their selves and associated mental states separately from its own. Most importantly, the agent may become aware of its own self by having multiple representations of it (together with their individual contents of awareness), that can "see" and control each other, if their current status allows them. These instances of the self and the associated mental states may refer to the past, as well as to possible future situations. Other possibilities include mental perspectives associated with assumptions, dreams, goals, etc., and, most importantly, metacognitive perspectives. Metacognition implemented based on this framework of mental states brings a new quality to the cognitive architecture. The agent in a metacognitively self-aware state (an instance of the agent's self labeled "I-Meta" in Figure 2 b) may be aware of its other instances of self, as well as mental attitudes attributed to them. This awareness means that other mental states and their components are reflected at a metacognitive level (and are accessible via "handles" provided by their representations) within the mental state "I-Meta". This metacognitive representation allows the agent to operate on its own mental states by changing their status and content. E.g., correction of a false awareness state (and associated with it plans that would make no sense to follow, given new information) may be possible to accomplish in a single flip of the status of the present state of awareness from "I-Now" to "I-False-Belief" and replacement of "I-Now" with another mental state developed in parallel in the background under a label "I-Surmise". E.g., the system under uncertainty may develop several parallel scenarios of future events and its own actions, keeping one of them as the *working scenario*. Upon acquisition of new critical information, the system may switch to another working scenario – this single-step operation may change behavior, future plans and interpretation of own actions in the past, without changing mental state contents at a cognitive level. In order to enable deliberate control of these abilities, the system needs to be explicitly "aware" of multiple instances of its self and of the attribution of mental attitudes to them.

Much greater capabilities become enabled in the system by metacognitive self-awareness in a long-term perspective of a cognitive growth. In fact, the kind of self-awareness described above allows the agent to deliberately re-design its own system of values and goals, regardless of the initially given goal (e.g., to make paper clips), using its personal experience and an acquired system of values.

A modern intelligent artifact is typically built for a certain specific purpose, which is, explicitly or implicitly, hard-coded in it "at birth". In contrast, human individuals have no life-long goals at the beginning of their lives. Development of a system of values and a specific goal in life may take a substantial part of the entire human life. In order to bring AI to a human level of intelligence in this sense, it will be necessary to implement in artifacts self-awareness and the ability to operate on their own values and goals, by means of an explicit attribution of values and mental states to multiple instances of the own self.

Illustration by Self-Regulated Learning (SRL)

Many concepts and recent state-of-the-art achievements in educational science have the potential to benefit AI. One example is SRL, which currently acquires increasingly

[1] When a cognitive state is explicitly attributed to an instance of a self, we say that it forms "a mental state". Specifically, by a mental state we refer to all cognitive states that are currently attributed to awareness of one instance of the self.

higher popularity in educational practices and only starts reaching awareness of the broad AI community. The notion of SRL includes a set of general metacognitive techniques, scaffoldings and strategies actively used by the learner to increase the learning efficiency (Zimmerman 2002). By its nature, SRL is essentially metacognitive (Winne 1995) and particularly relies on active self-awareness (Zimmerman 1995, 2002). Several schemes-architectures of SRL have been described previously (e.g., Matthews et al. 2000, Rheinberg et al. 2000, Winne and Hadwin 1998, Zimmerman 2002). Here we select a general three-cyclic SRL paradigm (Figure 3: based on Zimmerman 2002, Zimmerman and Kitsantas 2006) as an example to illustrate the leverage of metacognitive self-awareness.

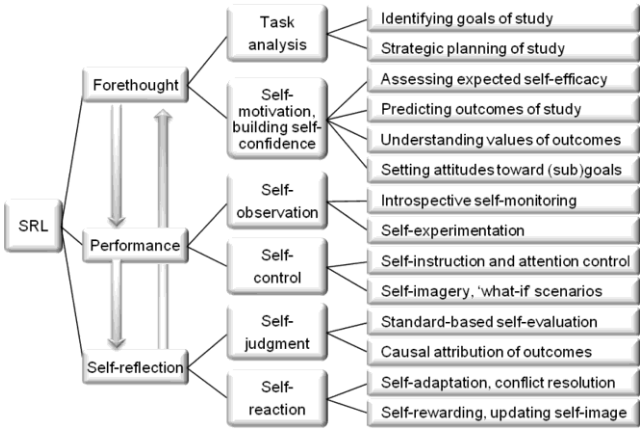

Figure 3. The cycle of three phases and the hierarchy of components of SRL (based on Zimmerman 2002, Zimmerman and Kitsantas 2006). Most components explicitly rely on the concept of self.

From Figure 3 we see that the selected general paradigm of SRL requires active manipulations with own mental states performed deliberately from a metacognitive perspective. Examples include selection of learning goals from the curriculum based on an estimate of own learning abilities, generating expectations of outcomes based on higher-level mental simulation of the learning process, experimentation with different strategies of learning in order to find causes of a failure, imagining self and instructing self how to behave in possible learning situations, metacognitive self-judgment, self-reward, self-adaptation, etc. In addition, the process of learning typically results in multiple alterations of the entire system of beliefs about the subject, which again suggests that self-control from a metacognitive perspective is necessary. These points are articulated with a specific example below.

In most cases, the failure of a student to make normally expected progress in learning can be attributed to a lack of self-regulation skills (Winne and Nesbit in press, Zimmerman 2002). Studies in education research show that learners that are not proficient in SRL make poor judgment of what they know, have difficulties in transferring their knowledge to new contexts, and do not always seek help usefully. The following example is taken from educational practice in the introductory college course in computer science, CS 112 at George Mason University. During one of the first recitation sessions, students in CS 112 learn the notion of the "*for*" loop, understanding its syntax and further technical details: e.g., how to use the "*if*" and the "*break*" instructions inside the loop. Subsequently, students receive an assignment to implement an algorithm that requires repeating the same sequence of operations 50 times. The code that a student turns in contains an explicit repetition of the sequence of instructions 50 times. When the teaching assistant asks the student: '*Can't this be done more easily by using a "for" loop?'* the student still cannot make a connection.

Such cases provide an opportunity for a computer-based intervention that will help the student to understand the connection, while simultaneously demonstrating the power of SRL. While computer-based learning environments (CBLE) that include intelligent pedagogical agents prove to be efficient in education at the cognitive domain level, their success in the SRL domain still needs to be demonstrated (Winne and Nesbit in press). The question is whether CBLE technology can be equally effective in fostering SRL skills, and if yes, then how this can be achieved. An example of a possibility follows.

Consider the following problem that could be assigned to students after the "*for*" loop concept was introduced: *"Write a Python code that will print out all the prime numbers less than N"*. In order to solve this problem, a self-regulated learner needs to set proximal goals as to how to approach it. A self regulating student may set, for example, the following proximal goals: (a) I want to write out what I know about prime numbers that can be useful here: their definition, properties, etc. (b) I want to design a plan and a procedure of figuring out how to determine whether a given integer is a prime number. (c) I want to identify relevant Python primitives and map the procedure onto a Python code. (d) I will monitor myself and evaluate my progress; if I detect that I am failing to follow the plan, then I will decide to look at other examples or seek help.

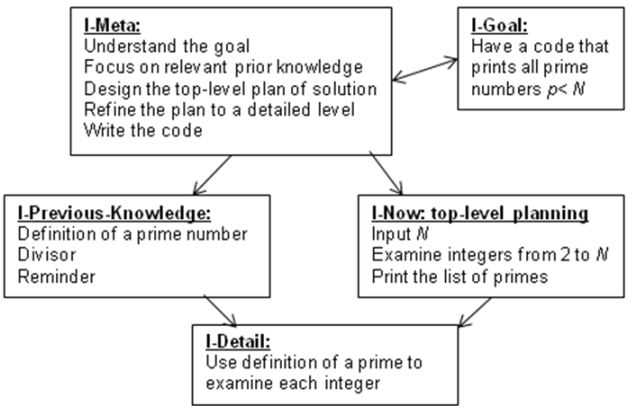

Figure 4. Problem solving using the general SRL paradigm and the mental state framework. Mental states (boxes) have self-explanatory labels (the underlined top line) specifying an instance of the self. Arrows indicate interactions among mental states.

These steps can be implemented in an intelligent tutoring agent at the metacognitive level, using the mental state framework, as illustrated in Figure 4. In order to help a student to successfully implement the solution and to acquire SRL skills at the same time, the agent will first perform a task analysis to determine the processes, steps, or procedure associated with performing this task, using illustrative models represented graphically on the screen. The set of interconnected mental states (Figure 4) will be created in working memory, but not displayed on the screen: these representations will be used to guide behavior of the agent. By mapping these mental states onto the context of the task and by processing them one-by-one, the agent will (a) remind the student of relevant prior knowledge – in this case, the definition of a prime number as "an integer greater than 1 that has only 1 and itself as its natural divisors", also the related notions of a divisor and a remainder; (b) sketch a plan of solving the task, starting from the top level and using built-in planning capabilities when elaborating details, (c) invoke and show on the screen the relevant schemas of Python primitives: e.g., the "*for*" loop schema used for repeated procedures, and finally (d) ask the student to use these blocks to produce a code that solves the problem.

In this scenario, the student and the agent equipped with a natural language interface will be able to interact as peers, cooperatively making progress toward their common goal. The key intrinsic element enabling this mode of operation in the agent is, again, the explicit representation of the self. Another necessary element is a notion of a schema (Samsonovich and De Jong 2005) that allows one to represent concepts and skills in one universal format.

Concluding Remarks

The example paradigm described above illustrated the leverage of self-awareness in learning how to apply prior knowledge to new problems. The SRL-based approach relying on self-awareness understood in the sense explained above results in increased robustness, flexibility and integrity of the cognitive process of practical learning.

One problem with transferring the available SRL techniques borrowed from educational science to the computational level of AI is that research in educational science is done at a relatively abstract, human-oriented, functionalist level, whereas any computational implementation requires specification of details and mechanisms. The key part of the challenge is to initiate and to bootstrap the transformation of knowledge from educational science to AI. The concept of self-awareness described above is particularly useful for the demonstration of key principles.

In cinclusion, when the cognitive system becomes explicitly aware of its own self, it gains the capacity to control its self-concept together with the attribution of beliefs and values to the self. The leverage of this mechanism is vital for paradigms that require self-management, one of which is SRL.

References

Cox, M. T., and Raja, A. 2007. Metareasoning: A Manifesto. *Technical Report BBN TM-2028*, BBN Technologies. www.mcox.org/Metareasoning/Manifesto

Cox, M. T., and Ram, A. 1999. Introspective multistrategy learning: On the construction of learning strategies. *Artificial Intelligence*, 112, 1-55.

Matthews, G., Schwean, V. L., Campbell, S. E., Saklofske, D. H., and Mohamed, A. A. R., 2000. In Boekaerts, M., Pintrich, P. R., and Zeidner, M. *Handbook of Self-Regulation*, pp. 171-207. San Diego, CA: Academic Press.

Raja, A., and Lesser, V. 2007. A framework for meta-level control in multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 15 (2): 147-196.

Rheinberg, F., Vollmeyer, R., and Rollett, W., 2000. In Boekaerts, M., Pintrich, P. R., and Zeidner, M. *Handbook of Self-Regulation*, pp. 503-529. San Diego, CA: Academic Press.

Russell, S. J., and Wefald, E. 1991. Principles of metareasoning. *Artificial Intelligence*, 49: 361-395.

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

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

Singh, P. 2005. EM-ONE: *An Architecture for Reflective Commonsense Thinking*. Ph.D. dissertation. Department of Electrical Engineering and Computer Science. MIT: Boston, MA.

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

Winne, P. H. 1995. Inherent details in self-regulated learning. *Educational Psychologist* 30 (4): 173-187.

Winne, P. H., and Hadwin, A. F. 1998. Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, and A. C. Graesser (Eds.). *Metacognition in Educational Theory and Practice*, pp. 277-304. Mahwah, NJ: Lawrence Erlbaum Associates.

Winne, P. H., and Nesbit, J. C., in press. Supporting self-regulated learning with cognitive tools. In Hacker, D. J., Dunlosky, J., Graesser, A. C. (Eds.). *Handbook of Metacognition in Education.* Mahwah, NJ: Lawrence Erlbaum Associates.

Zimmerman, B. J. 1995. Self-regulation involves more than metacognition: A social cognitive perspective. *Educational Psychologist* 30 (4): 217-221.

Zimmerman, B. J. 2002. Becoming a self-regulated learner: An overview. *Theory into Practice* 41 (2): 64-70.

Zimmerman, B. J., and Kitsantas, A. 2006. The hidden dimension of personal competence: Self-regulated learning and practice. In Elliot, A. J., and Dweck, C. S. (Eds.). *Handbook of Competence and Motivation*, pp. 509-526. New York: The Guilford Press.